CREATING AN OPEN GEODEMOGRAPHIC CLASSIFICATION USING THE UK CENSUS OF THE POPULATION

Christopher George Gale

Department of Geography

University College London

A thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy (PhD)

May 2014

Word Count: 88,482

Declaration

I, Christopher George Gale confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:

Date:

Acknowledgements

First and foremost, I would like to thank Professor Paul Longley, my primary supervisor, for his insight, time, guidance and most of all his patience throughout this PhD. I would also like to thank my secondary supervisor, Dr James Cheshire, for his encouragement; he has provided a calm voice of reason throughout. Paul and James's help has been invaluable and I am grateful for all the support they have both given me.

I am also grateful to Dr Alex Singleton for his input; his knowledge of the subject area meant it was beneficial to be able to discuss the research with him. Additional thanks goes to Dr Daniel Lewis and Dr Muhammad Adnan who at the time of starting this PhD were fellow postgraduate students, always on hand to offer advice and practical assistance.

I would like to acknowledge the Office for National Statistics, firstly for contributing towards the UCL Impact Award that funded this project and secondly for the input and support received from Andy Bates as their representative for this work. I am grateful for his time and help.

None of this would, however, have been possible without the ever-present support and understanding of my future wife, Katherine Penny. Katherine pushed me to keep going, even through the most challenging of times, and this PhD is testament to her love and encouragement.

Finally, I must thank my family and friends for putting up with me over the past three years. I am incredibly grateful to my parents for all that they have done for me and in particular for their enduring support, I would not have got this far without them. I would also like to thank my brothers and sister for tolerating their grumpy older brother and my future mother-in-law, Elizabeth Penny, for her constant supply of tea, biscuits and kindness whenever I worked at her house. All my friends have been so encouraging and I am grateful to them all; however special thanks has to go to Jill Hazleton and Holly Dolan who have gone above and beyond the call of duty to proof read my work; the time and effort they have put in, will not be forgotten.

Abstract

The 2011 Area Classification for Output Areas (2011 OAC) is a new open geodemographic classification of the UK based on 2011 UK Census data. The 2011 OAC, created in partnership with the Office for National Statistics (ONS), supersedes the 2001 Area Classification for Output Areas (2001 OAC) to provide the most current open geodemographic view of the UK.

The 2001 OAC was widely used in academia, local government and by commercial organisations, but its reliance on data from the 2001 UK Census has led to a perceived degradation of reliability over time and a decline in users. The release of the 2011 UK Census data provided the opportunity to create a 2011 OAC which could address some of the acknowledged flaws of the 2001 OAC, such as the methods used for data handling, to create a more robust methodology. The publication of this methodology with accompanying documentation, in addition to utilising open-source software, guarantees the reproducibility of the 2011 OAC; with an additional benefit of the methodology being able to act as a template for future bespoke open geodemographic classifications.

Open geodemographic classifications, unlike those provided by commercial organisations, have historically been unable to utilise ancillary data sources to enrich and update their systems. This research proposes an alternative approach; utilising the limited range of Open Data sources made available regularly at the small granular level to create uncertainty indicators. These indicators allow areas of uncertainty that develop over time within the classification's geodemographic assignment to be identified; allowing users the opportunity to take compensatory action.

This project delivered a new open geodemographic classification of the UK. The methodological advances, use of open source software and ability to assess the temporal stability of geodemographic assignments mean the 2011 OAC can be considered a step forward for open geodemographics.

Table of Contents

List of Figures

Chapter 2

Chapter 3

Chapter 5

Chapter 6

Chapter 7

Chapter 8

Chapter 9

Appendix C

Appendix D

List of Tables

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 8

Appendix B

Appendix C

Appendix D

List of Abbreviations

Abbreviation	Definition
2001 LOAC	2001 London Output Area Classification
2001 OAC	2001 Area Classification for Output Areas or 2001 Output Area Classification
2011 EW OAC	Preliminary 2011 England and Wales Area Classification for Output Areas
2011 LOAC	2011 London Output Area Classification
2011 OAC	2011 Area Classification for Output Areas or 2011 Output Area Classification
Acorn	A Classification of Residential Neighbourhoods
AMOEBA	A Multidirectional Optimal Ecotope-Based Algorithm
AZP	Automated Zoning Procedure
BCSS	Between-Cluster Sum of Squares
CCS	Census Coverage Survey
CSV	Comma-Separated Values
DUG	Demographics User Group
DWP	Department for Work and Pensions
DZ	Data Zone
ED	Enumeration District
ESRC	Economic and Social Research Council
GDP	Gross Domestic Product
GIS	Geographical Information Science/System
GLA	Greater London Authority
GROS	General Register Office for Scotland
IHS	Inverse Hyperbolic Sine
IMD	Index of Multiple Deprivation
IPS	International Passenger Survey
LFS	Labour Force Survey
Log	Logarithm
LSOA	Lower Layer Super Output Area

Abbreviation	Definition
MAUP	Modifiable Areal Unit Problem
MPS	Metropolitan Police Service
MSOA	Middle Layer Super Output Area
MYE	Mid-Year Population Estimates
NISRA	Northern Ireland Statistics and Research Agency
NRS	National Records of Scotland
OA	Output Area
OGL	Open Government Licence
ONS	Office for National Statistics
OPCS	Office for Population, Censuses and Surveys
OS	Ordnance Survey
OSM	OpenStreetMap
PAM	Partitioning Around Medoids
PAS	Public Attitude Survey
PC	Personal Computer
PCA	Principal Components Analysis
PiN	Pinpoint Identification Neighbourhoods
QNHS	Irish Quarterly Household Survey
RUC2011	2011 Rural-Urban Classification
SA	Small Area
SED	Squared Euclidean Distance
SIR	Standardised Illness Ratio
SOA	Super Output Area
TfL	Transport for London
UCL	University College London
UK	United Kingdom of Great Britain and Northern Ireland
USOA	Upper Layer Super Output Area
VOA	Valuation Office Agency
WCSS	Within-Cluster Sum of Squares

Chapter 1

Introduction: Aims and Structure

1.1. Introduction

The population of the UK is a large and complex entity that consists of millions of individuals, each with their own unique characteristics. Trying to comprehend the diversity of the UK population is a complicated task. Multiple data sources exist that detail different aspects of the UK's population, with a continual stream of information being made available from Government data repositories, such as data.gov.uk, which now contain hundreds of datasets on the theme of society. This constant flow of information from multiple sources makes it difficult to form any consensus on what the attributes of the UK's population are at finest levels of granularity.

The majority of datasets available through websites like data.gov.uk provide insufficient detail to offer a geographically in depth representation of the UK's population. The decennial UK Census provides an alternative to this, and is designed as a means of counting individuals in order to quantify their characteristics. Despite the UK Census only taking place every 10 years, it continues to provide the most comprehensive set of statistics relating to the country's population, both in terms of the range of questions asked and geographical scope. The Census forms are completed for all individuals in every household in the UK on a specific day, with the most recent occurring on the 27th March 2011. The data derived from these forms are not however released at the household or individual level for 100 years due to the laws governing confidentiality (ONS, 2012a). The release of data from the last two UK Censuses, held in 2001 and 2011, have instead aggregated individual and household data to custom areal units. The smallest of these units divided the UK into 232,296 areas in 2011 and contained on average 272 people each.

The volume of data released from the last UK Census in 2011 – hundreds of tables and thousands of variables – provides a detailed overview of the UK's population. The

availability of this amount of data for each of these units means that a lot of information is known about very small areas of the UK. The extent to which this array of data can be easily interpreted and summarised in an easily understandable format is limited. The level of detail offered by the latest UK Census makes it difficult to know which parts of it are relevant to any particular task. It is therefore necessary to process the raw data; in other words the data needs to be summarised, analysed or otherwise converted into usable information. That information can be more readily understood and used to provide summaries of the UK population.

One technique that can be utilised to generalise multidimensional datasets is the creation of area classifications. Instead of dealing with individuals, they classify areas into groups based on the similarities of certain properties within them. Area classifications can be utilised for a range of applications, from classifying forestland (Azuma and Monleon, 2011) to hazardous areas (Cox et al., 1990), providing a convenient simplification of a large dataset while also incorporating a spatial component. The application of this concept on a large multivariate dataset like the UK Census provides a method of simplifying the population and the variations that occur across the country. In this context an area classification forms the core component of geodemographics, which is "the analysis of people by where they live" (Sleight, 2004, p. 16).

Geodemographics can be considered the formalisation of the relationship between people and place. There are two core components: social similarity, independent of locational proximity and the spatial autocorrelation of like-minded individuals. The spatial autocorrelation embodies Tobler's First Law of Geography that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, p. 236); meaning people that live in the same area are more likely to share similar characteristics than those living further away. Although this is tempered by the existence of social similarities that are independent of location, acknowledging comparable population groups can be found irrespective of geographic location. A geodemographic classification utilising these principles therefore provides insight into the characteristics of the population based on where they live.

The ability of a geodemographic classification to distil large multivariate datasets into succinct descriptions of the population has led to their continued use since being developed by Richard Webber in the 1970's (Gale and Longley, 2013). In the preceding 40 years geodemographic classifications have been predominantly made available by

commercial companies. These systems have dominated the marketplace due to their use of ancillary data sources to regularly update their products, something that is popular with users. In terms of market share this has given them a competitive advantage over the limited number of freely available geodemographic classifications. The last open geodemographic classification created on a UK scale and made freely available was the 2001 Area Classification for Output Areas (2001 OAC). This was released by the Office for National Statistics (ONS), the body responsible for official Government statistics in England and Wales, in collaboration with Leeds University.

The creation of the 2001 OAC, based on only 2001 UK Census data, was a significant milestone in the development of a geodemographic system with a completely open, transparent and scientifically reproducible methodology (Vickers and Rees, 2007). Its widespread adoption meant the release of the 2011 UK Census acted as a catalyst for an updated version of the classification to be created. The creation of this classification forms the basis for this thesis.

Like the 2001 OAC, the creation of the new classification was in collaboration with the ONS. Funding from a UCL Impact Award, with a contribution from the ONS, meant the project began in late 2010, several months before the 2011 'Census Day'. The timescales of the project meant that for the first 28 months of funding no appropriate data from the 2011 UK Census were available. Although the timing of the project could be considered problematic, it provided an opportunity to explore other aspects of geodemographics. This allowed a more in depth understanding of key processes and data availability to be developed before they needed to be applied to creating a new classification – something that would not have been possible if the relevant data had been available at the start of the project.

A key aspect explored within the research project is temporal stability; which is an issue that exists with any geodemographic classification. A primary reason why commercial systems use non-Census data sources is so they can be updated regularly. In comparison, the 2001 OAC still reflects the UK's characteristics from 2001, as it would have needed another Census dataset to update it. This limitation has led to the 2001 OAC being considered out-dated by users. The perceived need to continually 'refresh' commercial systems to reflect changes in the UK is not always warranted. Sleight (2004) suggests change is inconsequential as certain types of people will always dominate certain areas and as people move out likeminded people replace them. Longley et al. (2011) also

indicates stability across a large proportion of the British population over hundreds of years.

The perception that the entire 2001 OAC now provides inaccurate representations of the UK's population characteristics due to changes that have occurred over the intervening years is fallacious. The future proofing of geodemographic classifications that cannot be 'refreshed' with regularly updated data sources was therefore a concept explored in the project. The investigation of the temporal uncertainty of the 2001 OAC's geodemographic assignment provided a good opportunity at the start of the project to explore the on-going relevance of the classification. It also provided an opportunity to become acquainted with datasets available to create an open geodemographic classification, and explore the role alternative datasets to the UK Census could be used and incorporated. From a practical perspective it also allowed the development of skills, such as learning to code in the open source R program (R Development Core Team, 2011) to perform a number of statistical and graphical operations.

The analysis of the temporal uncertainty of the 2001 OAC provided three key opportunities: it allowed an assessment of the current data environment in the UK to be carried out; it provided an opportunity to develop a method for assessing the stability of a geodemographic classification over its lifetime and it allowed the development of a skill base. These three components were all relevant to creating a new geodemographic classification following the release of the 2011 UK Census. Each element had an impact on the methodological approach taken and the decision making process.

Complicating the creation of a new UK-wide geodemographic classification was the staggered release of the 2011 UK Census data for England and Wales, Northern Ireland and Scotland at the smallest geographic levels. Data appropriate for use with a geodemographic classification were released in January 2013 for England and Wales, between January and February 2013 for Northern Ireland and December 2013 for Scotland. The aspirations of the ONS to release a new classification by July 2014 meant a UK-wide dataset needed to be processed as quickly as possible. The wait for the Scottish data and the benefits from performing the temporal uncertainty analysis however meant a methodology was already in place and had been extensively tested. The existence of a robust methodology meant the creation of a new geodemographic classification for the UK could be accelerated following the release of data for Scotland, and therefore the final classification could be produced with relative ease.

The aims of the thesis, outlined in Section 1.2, were designed to avoid the exact re-creation of the 2001 OAC methodology using 2011 UK Census data. To have done so would have been a wasted opportunity to advance open geodemographics, additionally it would not have guaranteed the best representation of the UK. The 2001 OAC was not however ignored, and was instead used as a guideline to illustrate the steps required to create an open geodemographic classification. The creation of a new geodemographic classification needed to acknowledge the heritage of the discipline, but this did not mean it had to be constrained by decisions made in the past.

1.2. Aims

The release of the 2011 UK Census data provided a catalyst for the creation of new and updated geodemographic classifications, both commercial and non-commercial. The principal aim of the project was therefore to use this opportunity to create a new open geodemographic classification of the UK at the smallest areal unit level. To achieve this a set of more specific secondary aims were required to give the project the necessary focus. These aims were derived from the literature review and background research undertaken in Chapter 2 and can be summarised as follows:

i) To create a new open, transparent and reproducible methodology.

The methodology of the 2001 OAC was designed to be reproducible (Vickers and Rees, 2007). This did not however mean that the processes used could not be improved upon. The reliance on commercial software, such as the SPSS statistical package, to create the 2001 OAC prevents the entire methodology from being considered completely reproducible or open. The new classification was therefore designed to either use open-source programs or be compatible with them.

There was also an opportunity to explore the methodological components of creating a geodemographic classification in greater detail. The continuing improvement in computational power meant certain aspects, such as data processing techniques, could be explored in greater depth (when compared to the 2001 OAC) to ensure the final classification was both robust and optimal.

ii) To consult with users to determine what their requirements are for the classification.

A new classification of the UK needed to reflect the requirements of users. The creation of the 2001 OAC showed that a classification of the UK at the finest spatial level was possible. The new classification therefore had a greater focus on users' desired outcomes. The emphasis on designing a classification to be conscious of the requirements of the user base therefore enhances the prospects of its wide scale adoption.

iii) To develop visual and descriptive outputs to facilitate users understanding of the results produced by the classification.

An essential part of any geodemographic classification is the production of a number of key outputs – these help a user to better understand the results of the population summaries produced. Cluster names and descriptions were provided as part of the 2001 OAC, supplemented with static maps as PDF files. The new classification of the UK needed to at least replicate these outputs, however the numerous developments in web based mapping and spatial data infrastructures (see Goodchild, 2007; Haklay et al., 2008; Singleton and Longley, 2009a; Carpenter and Watts, 2013) provided an opportunity to produce more engaging outputs.

The production of outputs for the new classification of the UK therefore needed to meet the minimum expectations of users, while incorporating the latest advances in geodemographics and geographic information science (GIS), to provide a suite of descriptive and visual tools for interpretation of the final results.

iv) To validate the classification once complete to assess the final outcome.

Validation was required to assess the overall quality of the final classification across a number of different categories. There are multiple components of a geodemographic classification that impact its robustness. These components therefore required evaluation to ensure the final classification provided the best representation of the UK.

v) To explore alternatives to using ancillary data sources to update geodemographic classifications that highlight the temporal stability, or otherwise, of resident populations.

The applicability of geodemographic classifications in the subsequent time periods that follow their release is of interest to users. Analysing the temporal stability of a geodemographic classification provides a method to highlight only the areas of the country that experience high levels of change over a set period of time. Investigating the

use of temporal uncertainty indicators is therefore advantageous, as, unlike commercial systems, they do not require the wholesale modification of the classification with ancillary data. Instead they can highlight areas in the UK where the geodemographic assignments may have become less stable over time and therefore require further attention.

1.3. Thesis Structure

This thesis is divided into nine chapters. Chapter 1 provides an introduction to the project and details the specific aims. Chapters 2 and 3 provide background information relevant to the project. Chapter 4 introduces the core concepts of the new UK-wide geodemographic classification. Chapter 5 is the first analytical chapter, and provides an overview of the investigation into using temporal uncertainty indicators. Chapter 6 provides the methodology of the new classification, while Chapter 7 details its implementation. Chapter 8 validates several aspects of the classification and Chapter 9 summarises the findings of the project and potential directions for future research. Detailed summaries for Chapters 2 to 9 are outlined below.

1.3.1. Geodemographics and Area Classification

Chapter 2 introduces the theories, principles and practices of area classifications and geodemographics. It provides an overview of the history of geodemographics and looks at the current geodemographic systems from both a practical and ideological perspective, exploring the differences between current geodemographic classifications. The previous national geodemographic classification built from the 2001 UK Census is discussed (the 2001 Area Classification for Output Areas or 2001 OAC), with particular focus on the methodology that underpins it. This is followed by an examination of the challenges associated with classifying London alongside the rest of the UK given the diversity of its population along with other potential pitfalls found within geodemographics. The purpose of his chapter is to provide a historical overview of geodemographics and area classification to allow the research agenda to be formulated.

1.3.2. The Census and Open Data

Chapter 3 introduces the UK Census and Open Data as data sources that can be utilised within a geodemographic classification. The history of the UK Census is detailed, and an

appraisal of its use as a data source in term in terms of scope, population coverage and data quality is made. The geographies that are used to disseminate the data are discussed, and how variations exist between different countries in the UK. In addition, the future of the UK Census is examined, along with how the prospect of changes to small area data provision in the UK may impact geodemographic applications. This leads into an exploration of the future of Open Data in the UK, and the role such data sources could have in current and future geodemographic classifications. The purpose of this chapter is to provide an overview of the UK's current data environment and outline which sources can be used in the creation of a new geodemographic classification.

1.3.3. A New Area Classification

Chapter 4 introduces the 2011 Area Classification for Output Areas (2011 OAC) and outlines the key concepts that influenced its construction. The key concepts are explored further, with focus on the influence of internal factors, such as the desire for the classification to be open, and external factors, like the availability of data. This is followed by the detailing of procedures used to engage with potential users of the 2011 OAC and the influence that these results had on creating the new classification. The key points from the user engagement results are explored and distilled into a set of user requirements. In addition, the steps required to incorporate these requirements into the construction of the 2011 OAC are also described. The purpose of this chapter is to identify the key concepts and user requirements for the new classification, and to outline how the design criteria and implementation of the 2011 OAC was shaped as consequence.

1.3.4. Temporal and Spatial Stability of Small Area Classifications

Chapter 5 examines the spatial and temporal uncertainty of geodemographic classifications. It explores how commercial classification systems use ancillary data sources to regularly update their products, something not possible with geodemographic classifications built with only Census data. An alternative to the commercial sector methods examined in this chapter is the creation of spatio-temporal uncertainty indicators from the limited amount of appropriate data available. The concept of uncertainty indicators being based on the premise that change in the UK varies both spatially and temporally is discussed, and how this can be used to provide a general

indication of how different areas exhibit different change characteristics. The implementation of these indicators is explored using the example of the 2001 OAC, with the benefits of being able to identify areas of significant change in the UK compared to having to regularly update a geodemographic classification with updated data being assessed. In addition, limitations of temporal uncertainty indicators are examined, and how the release of additional small area data sources could help address these issues. The purpose of the chapter is to propose and test a method that can be used as a viable alternative to the traditional approach for updating geodemographic classifications; and to provide an understanding of the advantages and limitations that exist with the techniques proposed.

1.3.5. Methodology for the 2011 Area Classification for Output Areas

Chapter 6 outlines the methodology used to create the 2011 OAC. The role of cluster analysis in constructing the 2011 OAC is explained along with specific details on the individual steps required to create the classification. The processes involved in identifying and listing variables and then reducing this list to make a final selection are explored in further detail. The different techniques used to prepare the data for clustering are discussed, as are the criteria used to select an optimum combination of techniques to create the 2011 OAC. Additionally, the clustering processes used to create the structure of the 2011 OAC is explained, along with how the different types of outputs produced help give names and descriptions to the groups created. The purpose of this chapter is to provide a detailed methodological overview of the different techniques used to create the 2011 OAC and explain why decisions were made. Chapter 7 details the outputs created using the processes described in this chapter.

1.3.6. Creating the 2011 Area Classification for Output Areas

Chapter 7 outlines how the methodology discussed in Chapter 6 was implemented to create the 2011 OAC. Discussed are the inputs that were used to create the 2011 OAC and how the final selection was made; the processes used in the three methodological stages of initial variable identification, variable reduction and final variable selection are also detailed. The outputs from the different data processing techniques applied to the final variable selection and the outputs of these processes are examined in order to identify the optimum dataset used as the basis of the 2011 OAC. In addition, the results

cluster analysis are discussed and how the core outputs of this process can be split into two categories: descriptive and visual. A number of additional outputs from the 2011 OAC are also detailed, such as the release of the R code used to construct the classification, to fulfil the aspiration for all processes undertaken to be as open as possible. The purpose of this chapter is to provide evidence of the successful implementation of the methodology outlined in Chapter 6, and to discuss the number of different outputs of the 2011 OAC.

1.3.7. Validation of the 2011 Area Classification for Output Areas

Chapter 8 details different validation exercises performed on the 2011 OAC. Discussed are the different categories used to validate the classification: variable performance; cluster assignment certainty; homogeneity; changes between 2001 and 2011 and ground-truthing. Each of the categories is explored in detail, with focus on how effective the 2011 OAC is at providing accurate representations of the UK's population. Additionally, the categories analysed cover several different aspects of the classification's construction and outputs, therefore allowing multiple attributes of the 2011 OAC to be assessed and discussed. The purpose of this chapter is therefore to provide evidence of the overall robustness of the 2011 OAC in how it represents the diverse characteristics of the UK's population.

1.3.8. Conclusions and Future Work

Chapter 9 summarises the research undertaken to create the 2011 OAC. It explains why the overall aim of the project, to create a new open geodemographic classification of the UK, can be judged a success. In addition, this chapter looks at the contribution made by each of the secondary aims; followed by an exploration into what some of the possible applications of the 2011 OAC will be and how the current economic climate of the UK may impact its use. The lifespan of the 2011 OAC is also discussed, and how as the classification ages what role academic research could have in keeping it updated and relevant to the changing population dynamics in the UK. The chapter concludes by summarising the 2011 OAC and where it sits in the current UK geodemographic classification market.

Chapter 2

Geodemographics and Area Classification

2.1. Introduction

This chapter introduces the theories, principles and practices of area classifications and geodemographics. Section 2.2 defines what area classifications and geodemographics are and how they relate to both scientific and statistical theory. Section 2.3 gives an overview of the history of geodemographics from its origins over 100 years ago. Section 2.4 looks at the current geodemographic systems from both a practical and ideological perspective, exploring the differences between current geodemographic classifications and examining potential future developments. Section 2.6 details the previous national geodemographic classification built from the 2001 UK Census. It gives an overview of the structure of the classification and the methodology that underpins it. Section 2.7 discusses the challenges associated with classifying London alongside the rest of the UK given the diversity of its population. Section 2.8 looks at the potential pitfalls found within geodemographics, and how much these problems relate to modern iterations of classifications. Finally, Section 2.9 draws these points together to form a research agenda for this project to follow.

2.2. Area Classification and Geodemographics

Area classifications seek to group residential areas together based on their similarities. Webber and Craig (1978) suggest that area classifications provide a unique way of viewing patterns formed from multiple variables that differ from one area to the next. The concept of sorting and categorising things based on similarity is not a new one and is just an extension of the human condition to simplify the world around us. The population of UK was 63.2 million in 2011. These 63.2 million individuals can all be considered to be unique, yet share enough characteristics to be grouped together into

similar categories based on social structure, economic conditions or cultural behaviour values.

Area classifications can be divided into two types: categorical and continuous. Categorical classifications assign areas to a particular group, for example the 2011 rural-urban classification (RUC2011), which assigns areas in England and Wales to one of four urban or six rural categories. Continuous classifications assign areas a value, and where each area falls on the spectrum determines the classification outcome. An example of this is the English Indices of Deprivation 2010. This uses 38 separate indicators to calculate the Index of Multiple Deprivation (IMD) (Department for Communities and Local Government, 2011a), a de facto way of classifying the interactions between poor physical and social conditions across England. Indicators are used to calculate a deprivation score on a continuous scale. These values can be grouped together and ranked so that areas with the most deprivation can be identified.

Although ranking yields ordinal data, it is not uncommon for these results to be treated as continuous data when there are five or more categories (Johnson and Creech, 1983; Zumbo and Zimmerman, 1993). Figure 2.1 is the IMD rescaled for London, where the ranks have been split into deciles, grouping the population somewhere on the scale between the most to least deprived. Geographic groupings of areas experiencing greater or lesser deprivation clearly emerge, which is something that could not be easily identified by only looking at the raw data.

The simplification of complex relationships that are found within multivariate datasets, and an incorporation of the spatial elements that exist within area classifications form the core component of geodemographics. Geodemographics is “the analysis of people by where they live” (Sleight, 2004, p. 16). This is based on the concept that similar people are more likely to live within the same locality and that such area types will be distributed in different locations across a geographical space. Geodemographics can be seen as formalising the relationship between people and place. As such, knowledge of an individual’s home location can provide a great deal of insight into their identity (Vickers and Rees, 2007).

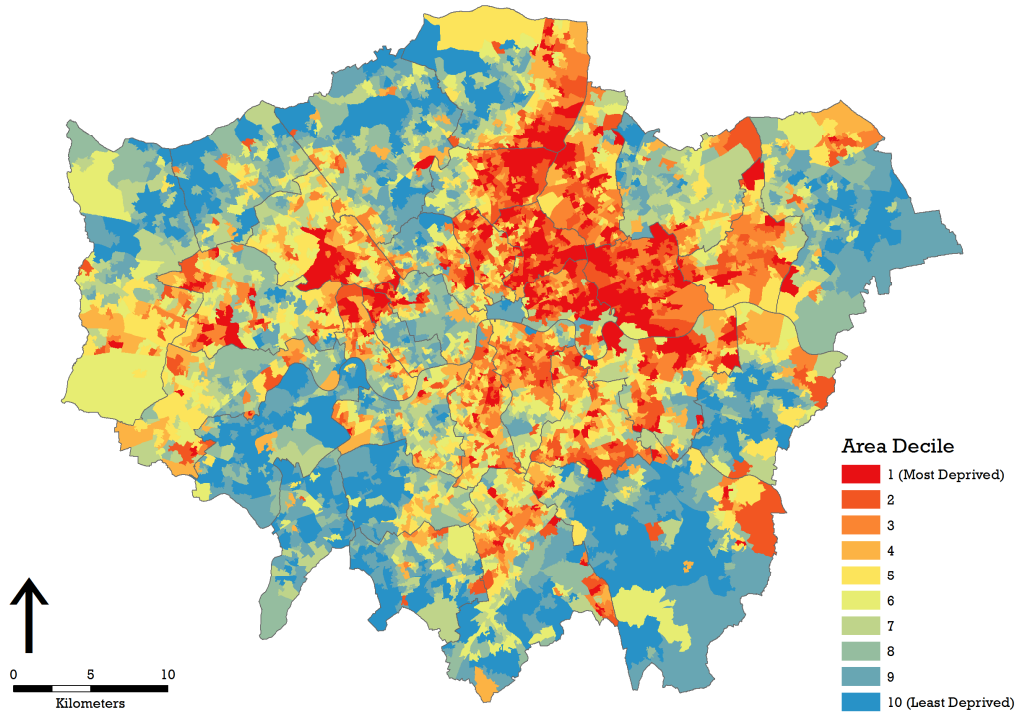


Figure 2.1: Overall Index of Multiple Deprivation for London in 2010

Department of Communities & Local Government, Indices of Deprivation 2010. © Crown Copyright and Database Right 2014.

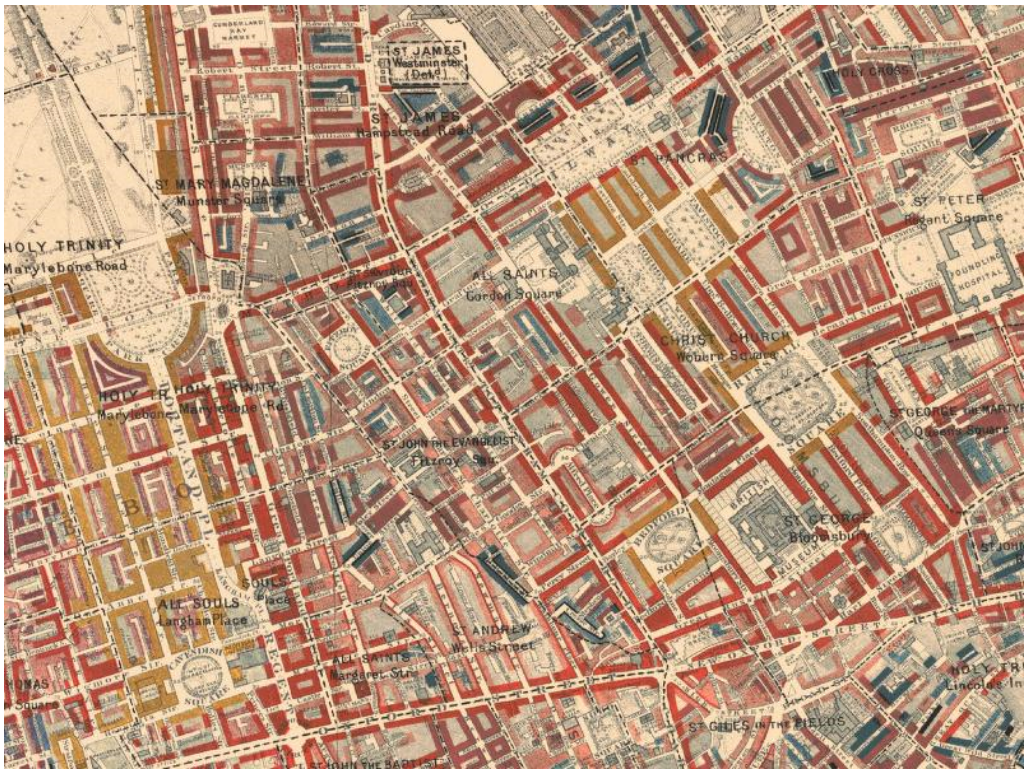
Geodemographics actively seeks to identify patterns in multidimensional datasets to group the population together by their various characteristics; these groupings provide an overview or 'picture' of the population within that neighbourhood. It is considered to be one of geodemographics key strengths that by knowing something about a neighbourhood you can infer information about an individual (Weiss, 2000; Sleight, 2004). The core components of geodemographics are social similarity, independent of locational proximity and the spatial autocorrelation of like-minded individuals. The spatial autocorrelation embodies Tobler's First Law of Geography that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, p. 236). Therefore two households located next to each other have an increased likelihood of sharing similar characteristics, both in terms of the physical attributes and the residents who live inside. A core function of geodemographics is the identification and grouping of areas that share similar characteristics but are not connected geographically.

An ever-increasingly mobile society would seem to conflict with Tobler's Law, with it having little relevance beyond a local scale for geodemographics. A mobile society now means that social structures are replicated in different, but not necessarily contiguous parts of the country, making Tobler's law insufficient by itself to fully explain population characteristics seen across a country. Vickers (2006), by adding a geodemographic element to Tobler's law, coined the first law of geodemographics as: "people who live in the same areas are more similar than those who live in a different area, but they may be just as similar to people in another area in a different place" (p. 16). This allows for geodemographic classifications to be built providing summary indicators of the social, economic and demographic characteristics of neighbourhoods (Adnan et al., 2010).

The aggregations of individual and household characteristics and neighbourhood attributes have been widely used for resource planning and allocation in both the commercial and public sectors (Shelton et al., 2006). Geodemographic classifications, are only as good as the data used to construct them. They contain no logic or advanced knowledge of the data they contain. They should not be used to answer a question, but rather to identify what questions should be asked. As such, they are an inductive process, or put more simply information should stem from the data and not the user (Openshaw, 1994). A geodemographic classification can therefore be seen as the practical output of the theory that underpins geodemographics as a research field and can trace its intellectual heritage back to the 19th century.

2.3. The History of Geodemographics

Geodemographics is considered by many to have been pioneered by Charles Booth in his studies of deprivation and poverty of London between 1898 and 1899. Published in 1899 as the Descriptive Map of London Poverty, the London-wide study combined direct observations of poverty and deprivation indicators with visits to households. Collating the data meant that each street in London could be described by the general socio-economic condition of its inhabitants (Harris et al., 2005). Figure 2.2 is an extract of a map that can be considered the world's first social area classification, showing the geographic variability of the seven different groups identified by Booth.



BLACK: Lowest class. Vicious, semi-criminal.

DARK BLUE: Very poor, casual. Chronic want.

LIGHT BLUE: Poor. 18s. to 21s. a week for a moderate family

PURPLE: Mixed. Some comfortable others poor

PINK: Fairly comfortable. Good ordinary earnings.

RED: Middle class. Well-to-do.

YELLOW: Upper-middle and Upper classes. Wealthy.

A combination of colours - as dark blue or black, or pink and red - indicates that the street contains a fair proportion of each of the classes represented by the respective colours.

Figure 2.2: Section of Charles Booth's 1898-1899 Poverty Map of London

Source: booth.lse.ac.uk

The shading used to delineate streets indicates the existence of both homogenous and heterogeneous groups, with streets exhibiting numerous characteristics combining the shades of multiple categories, in all but name creating a fuzzy classification (Harris et al., 2005). Booth's classification was based almost entirely on subjective data sources, yet remains a good indicator of London's social and economic landscape today with social class data from the 1991 UK Census indicating London has in the most part remained unchanged (Orford et al., 2002).

The impact of Booth's research can clearly be seen in subsequent work, such as that of Marr (1904) on the housing conditions in Manchester and Salford. Figure 2.3 is a section of a map produced that divides the study area into one of ten categories. The map has a similar cartographic approach to Booth, but with colouring of the housing blocks rather than the streets. The maps produced by Marr and Booth highlight an important part of modern day geodemographics – visualisation. Whilst it is possible to convey the results via traditional means of text and statistical tables, the use of visualisations provides an output that can be quickly and easily interpreted by most users.

While no doubt influential, the work of Booth, Marr and their contemporaries lacked the detailed coverage achieved in current geodemographics through the use of modern data sources. The intellectual heritage of quantitative geodemographic analysis can be traced back to the work on urban studies by human ecologists of the Chicago School in the 1920s and 1930s. The period between this and the 1970s are discussed in Singleton and Spielman (2013), where a large body of work focussing on social area analysis (Shevky and Williams, 1949) and factorial ecology was undertaken. These methods, a continuation of the work of the Chicago School, provide a framework for social measurement to be empirically undertaken to better understand neighbourhood characteristics (see Longley, 2005). The work of Shevky and Williams (1949) in Los Angeles, and that of Shevsky and Bell (1955) in San Francisco are important stepping stones towards what we know and understand modern geodemographics to be. For the first time they used solely statistical methods to classify areas based on their social composition. This desire to generalise urban social patterning in the 1970s led Richard Webber to develop a branch of applied urban studies that he would later term 'geodemographics' (Gale and Longley, 2013).

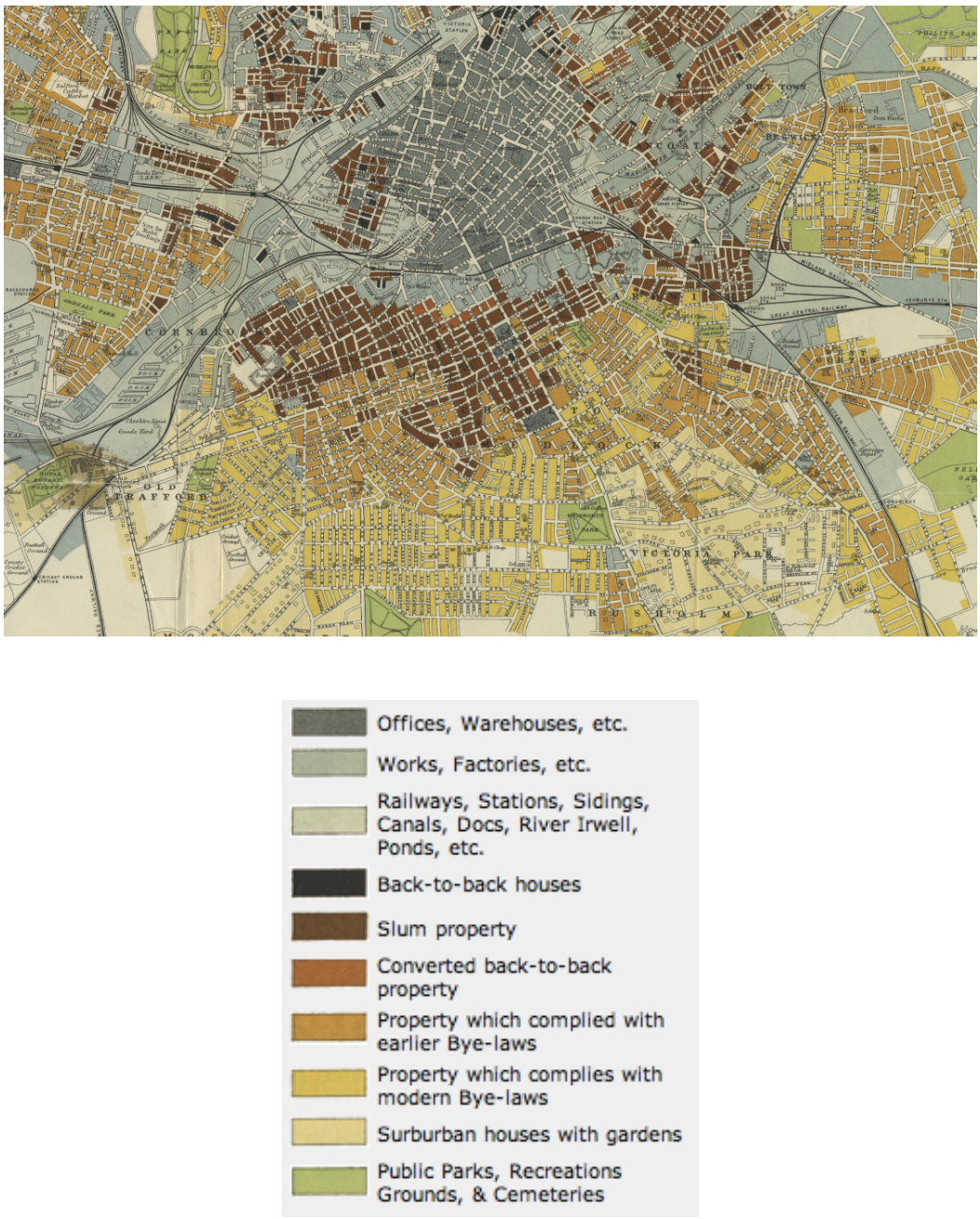


Figure 2.3: Section of Marr's 1904 Housing Conditions Map of Manchester

Source: manchester.publicprofiler.org/marr

Geodemographic classifications emerged as a methodological solution for handling highly dimensional Census data (Webber, 1978). Webber (1975) constructed a social area study to identify deprived inner city areas of Liverpool, before moving on to creating national level classifications (Webber and Craig, 1976, 1978; Webber, 1977). It was at this point that geodemographics started to dilute away from being a pure academic led field, with commercial interest being shown both in the UK and the USA. This trend continued through the 1980s, with geodemographics becoming a by-word for private sector marketing, with a large number of different geodemographic classifications being made available. In the UK the release of data from the 1981 Census spurred development in the commercial sector, and led to the divide that still exists today between commercial and academic geodemographics, namely the use of ancillary data sources. Prior to this period the sole data source for most geodemographic classifications had been national Censuses. The multitude of commercial companies trying to get a competitive advantage in a crowded market place meant an increased focus on providing regular updates to their classifications and the inclusion of data sources not part of a traditional Census. These data included income, the electoral register, vehicle registration data, county court judgement, credit reference agency data and lifestyle data (Sleight, 2004; Harris et al., 2005). The linking of Census data with these ancillary data sources allowed for significantly better discrimination of consumer behaviour (Batey and Brown, 1995), and classifications such as CACI's Acorn (A Classification of Residential Neighbourhoods), Experian's Mosaic, PiN (Pinpoint Identification Neighbourhoods) and Super Profiles came to the forefront.

The release of the 1981 UK Census was also significant for the academic side of geodemographics. It started the decennial trend of free geodemographic classifications after each new Census (see Charlton et al., 1985; Rousseeuw, 1987; Blake and Openshaw, 1994, 1995; Vickers and Rees, 2007). Similar trends continued during the 1990s and into the 21st century, with commercial geodemographics, at least in the mind-set of the users, leading the agenda – an outcome of their superior data resources, continuing use of more and more ancillary data sources and well-funded marketing departments. It was from this perceived commercial influence that criticisms of geodemographics emerged. Goss (1995) describes geodemographic classifications as being an over-simplification of society; noting that if social identity is both defined and sold by marketers then by extension it could be manipulated by them as well. This assumes that social identity is derived for consumer choices, something that Holt (1998) would seemingly disagree with, stating that consumer choices do not automatically equate to class reproduction.

The views of Goss (1995) are synonymous with an era of geodemographics that was seemingly fixated on providing a tool for marketers and marketing. In the decades since, geodemographics has moved beyond this narrow viewpoint, as can be seen by the current multitude of applications available.

2.4. Modern Geodemographics

Modern geodemographics is a wide and varied field. The current range of geodemographic classifications available present a range of options for users, although they are likely to come with a price tag. Singleton and Spielman (2013) identified a total of ten national geodemographic classifications available for the UK in 2012, with all but one of them being a commercial offering. A similar pattern is identified in the USA, with nine national level geodemographic classifications, with only two of them being freely available. These figures suggest geodemographics has undergone a significant shift away from its academic roots. The academic led discourse remains strong, as geodemographics is more than just a set of varying national level general-purpose classifications; indeed such classifications are perhaps the most critiqued element of the field. Openshaw (1983) stated that "there is no magic universal statistical test that can be applied nor is there any possibility of deriving a classification suitable for all purposes" (p. 245). Openshaw et al. (1980) compared two geodemographic classifications, one on a local scale the other national and found that the resulting representations from each were different.

Voas and Williamson (2001) take issue with another part of the process, that the variation between the groups in a geodemographic classification is smaller than the variation within each group. This would suggest that groups identified in such geodemographic classifications are too dissimilar to ever be considered useful. The solution to these criticisms is creating more bespoke geodemographic classifications that are driven by a particular task rather than a one-size fits all solution. This has now become a possibility thanks to increases in computation power and a fall in the associated costs. Singleton and Spielman (2013) assembled a list of publications from the UK and USA that applied some variation of a geodemographic classification and found that UK academics had over double the publications in comparison with their counterparts from the USA. The other interesting pattern to emerge is the way the classifications were used, with Figure 2.4 detailing a list of different themes. The extent to which these studies are just applying geodemographic classifications that already

exist, or are creating are their own bespoke classifications is unclear. What is clear is that geodemographic classifications are still widely used in academia, although a distinction between analytical studies of existing classifications, and creating new bespoke classifications has to be drawn.

As Figure 2.4 indicates, academic led geodemographics addresses a wide and varied number of topics. Although this discourse is significant, it is important to focus on users of such systems. They are more likely to be concerned with practical elements, rather than the number of papers published. It is likely the majority of non-commercial users choose these geodemographic classifications because they are free, rather than selecting them because they were constructed within an academic framework. At best this would be an added benefit, rather than a primary concern. These classifications remain popular with users, however there is only a limited number to choose between and they have received some criticism within the academic literature (Singleton and Spielman, 2013). The challenge is fitting the needs of users within an academic framework that seeks to push the geodemographic agenda forward. To that end Singleton and Spielman (2013) stated that "[t]he grand challenge for geodemographic systems is substantiating that they reflect real divisions in society, not chance grouping in the data" (para. 16). The increasing availability of data resources make fulfilling this criterion easier as they can be used to construct and evaluate classifications.

The historical focus of Census-based geodemographics in academia would suggest such systems are at a disadvantage when compared to the offerings of the commercial companies. To counteract this, a shift in the way academics approach creating geodemographic classifications is needed, although the continued reliance upon decennial Censuses means this is unlikely to happen. An event such as ceasing a traditional Census would provide an opportunity to develop and evaluate new methodological approaches to counteract the loss of such a valuable and reliable data source. The Census remains the most complete data source for any geodemographic classification, so will continue to underpin systems. This focus on Census data is likely to mean academic geodemographic classifications will continue to be less favourably compared with commercial alternatives. If, however, there are not going to be any future traditional Censuses (ONS, 2013a), then whatever the replacement is will impact how geodemographics as an academic research field functions in the future and could bring it more in line with how current commercial companies operate.

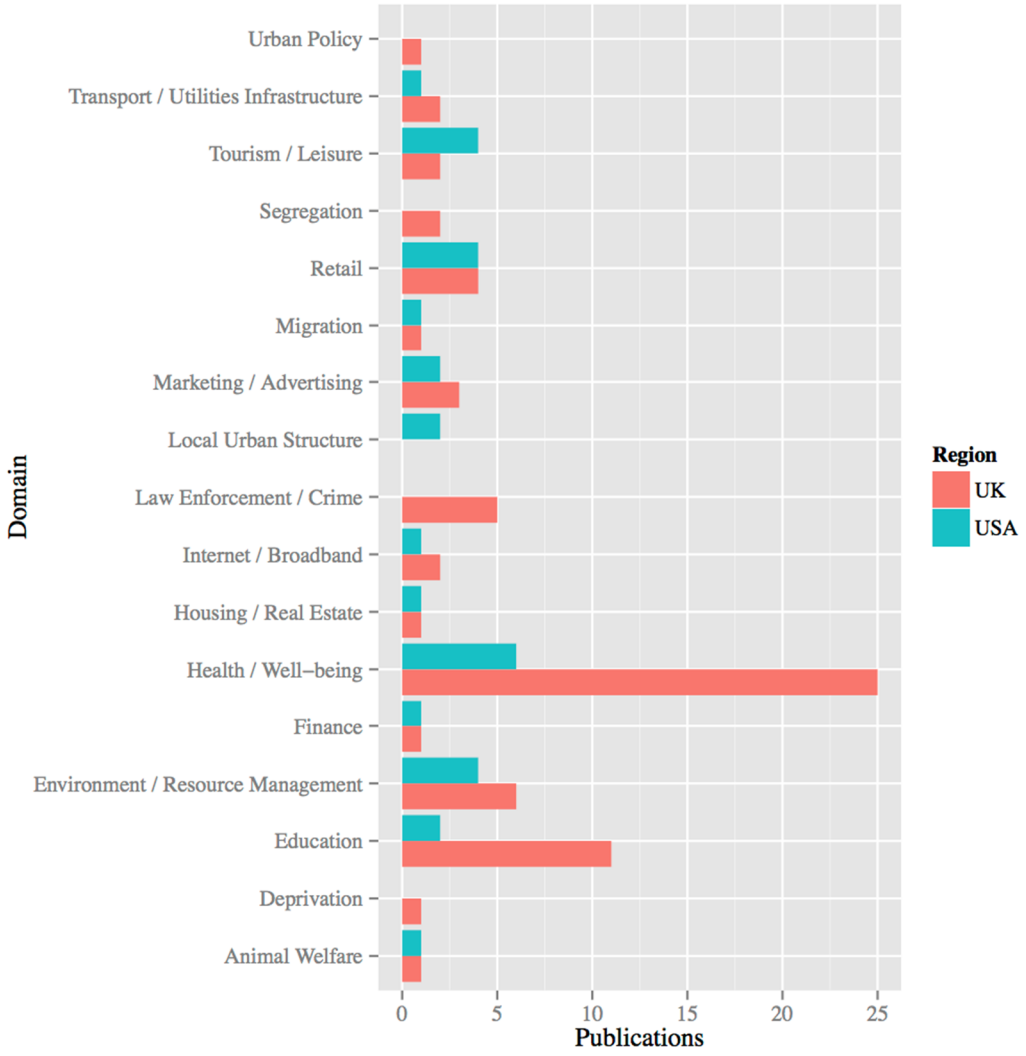


Figure 2.4: Geodemographic publications by theme and region

Source: Singleton and Spielman (2013)

2.5. Current Geodemographic Systems

Geodemographic classifications can be categorised into three main types. Commercial systems that either have a mass-market appeal or offer a specialised focused variant aimed at a particular market. Academic designed systems that are not necessarily as user focused and vary in size and scale depending on the research agenda. Finally, systems designed by third parties for internal use, such as local authorities who need bespoke classifications of their local area but are unable to pay for commercial products. The third category is most likely to go undocumented, so there is no meaningful way of knowing

how many of these exist. One exception to this has been created by Kingston upon Hull's city council (Feldman, 2011). The two other categories are easier to document and the current commercial systems available in the UK are shown in Table 2.1. This ease of identifying geodemographic classifications ultimately leads to comparisons between the commercial and academic systems available. This is a natural consequence of both types of product being available in the public domain, even if they differ significantly in their ideological approaches.

Progress in academic geodemographics is more disjointed than the seemingly smooth curve of progress in the commercial sector. Academic geodemographic research is peppered with well-meaning initiatives, but once the project ends and/or the funding ceases then a large number stop being supported. Outputs therefore do not tend to have a long shelf life, and any support offered for them is limited. The only exceptions to this are geodemographic classifications made for bodies such as the Office for National Statistics (ONS). The convoluted nature of funding for such projects means that the quantity and quality of the research taking place at any one time fluctuates, and this uncertainty can lead to academic geodemographics being too project focussed, and not placing enough importance on where it sits within the wider field. This makes progress challenging, as projects spend time going over the same ground, and only towards the end of the research process can the unique contributions be found. This is a systematic problem in any research field as broad as geodemographics, although the extent to which this is a real problem is debatable. The creative freedom that exists in academia, because research agendas are not driven by the need to make a profit, could in fact be considered an advantage that academic geodemographics has over the commercial offerings. While this results in fragmented discourse, it further progresses the research agenda in ways not possible in the commercial sector.

Table 2.1: Commercial geodemographic classifications available in the UK in 2013

Name and Company	Number of levels	Group numbers per level	Variables	Spatial level
P^2 People and Places, Beacon Dodsworth	3	14/41/157	Over 80	Smallest census geography
Mosaic, Experian	3	15/67/252	440	Unit postcode
Acorn, CACI	3	6/18/62	Over 400	Unit postcode
Cameo, Callcredit Information Group	2	10/57	Unknown	Unit postcode
Cloud Client, Cloud Client Ltd	1	15	29	Smallest census geography
Sonar, Redmoran	3	6/24/80	225	Postcode
Censation, Maw Data Solutions	3	5/19/53	600	Smallest census geography
Personicx Geo, Acxiom	1	60	Over 400	Postcode
Citizen, Marketing Metrix	2	6/28	Unknown	Postcode

Adapted from Singleton and Spielman (2013)

The approaches taken in commercial geodemographics are directly related to the internal forces of the commercial market they sit in. The main driver of current classifications such as CACI's Acorn or Experian's Mosaic come from the need to create a unique proprietary product within a crowded marketplace (see Table 2.1), leading to differences between how these commercial systems – both in the outputs produced and identifying their target audiences – are formed. The latest version of Acorn, released in early 2013, markets itself as the most up-to-date geodemographic classification that is also future proof (CACI, 2013a), as it no longer relies on decennial Census data. Indeed, the latest Acorn is a UK-wide classification that did not initially use any 2011 Census data at the smallest spatial level for Scotland because none existed in the public domain prior

to its release. This is perhaps the first example of a leading commercial geodemographic classification releasing a new version that does not use any Census data for a large proportion of its geographical coverage. There is no way to know this for sure as the technical documents containing the detailed methodologies used in the construction of commercial geodemographic classifications are not in the public domain, leading to Longley and Singleton (2009) to describe such approaches as "black-box". It is however testament to the progress within commercial geodemographics that within a 40-year period the Acorn classification has gone from being a brand attached to Webber's national Census ward level classification (Webber, 1977), to creating a geodemographic classification that in places does not need any Census data at all.

The extent to which methodologies of commercial geodemographic systems differ remain an unknown quantity. The 'black-box' approach means differences can only be evaluated by looking at how each classification is structured, and by examining the marketing materials. This lack of transparency forms the basis of many academic criticisms of commercial geodemographics; particularly as it prevents the reproduction of findings. The lack of validation makes it difficult to know if classifications form as accurate a representation of local neighbourhoods as described in marketing materials. It is possible that a higher level of precision could be implied that may not exist, thus overstating the capabilities of a classification. Due to a lack of transparency within the commercial systems there is reliance placed on the providers of such classifications that their outputs are correct. Without detailed explanation of how such conclusions are made users have to assume the processes are valid, and not just based on educated guesses.

However, despite the concerns expressed relating to validity and reproducibility, there are academics who have spoken out in defence of commercial geodemographics. For example Harris et al. (2005) state that in their experience commercial developers apply carefully thought out methodologies, adopt advanced methods and in some cases and have years of experience in classification development. Consequently, they argue the only real core difference between commercial and academic geodemographics is the latter's willingness to share their methodologies within the public domain (Harris et al., 2005).

Driven by commercial interests, there is no compelling reason why providers of commercial geodemographic systems need to release detailed information about the

methodologies or data sources they use. The commercial sector can be seen as a driver in the uptake of geodemographic systems. Users of these systems appear to place less importance on openness and transparency than those in academia, instead valuing timeliness of delivery, regular updating and increased relevance of the classifications to their own applications. It is unlikely commercial geodemographics will ever provide enough documentation to fully satisfy those who want to know the methodological underpinnings of their classifications. The role of academic geodemographics can therefore be seen to provide transparent methodologies that clearly explain the steps taken in creating geodemographic systems.

The recent focus in academia has been on small-scale bespoke geodemographic classifications, such as in policing (Ashby and Longley, 2005), education (Singleton, 2010) and local government (Longley and Singleton, 2009). What these have in common, with previous academic classifications, is they use data provided at a single spatial level (Charlton et al., 1985; Blake and Openshaw, 1994, 1995; Vickers and Rees, 2007). This differs from the commercial systems that use data that is provided at various spatial levels and fit it into their models and then project characteristics down to the unit postcode or household level (see Table 2.1). Undoubtedly the commercial approach provides a greater level of information at the smallest scales. It however also increases the level of uncertainty associated with each group assignment. This is coupled with the use of non-Census data sources, which may be easier to update but can lack the same statistical robustness of the Census. The extrapolation of data, in particular survey data, from a small cohort to give broader coverage introduces inherent uncertainties. These uncertainties can then be magnified when incorporated into a classification.

The approach of the commercial sector means academia is unlikely to create like-for-like alternatives to commercial geodemographic classifications. The range of data sources collated at different spatial levels and applied by these companies is vast, however little is known about their validity or spatial consistency.. The scope for creating anything similar is too large with little benefit. Instead of being shackled to legacy products or brands, academics can create variants based on the need to be open and transparent, while incorporating certain elements from commercial classifications. This is especially relevant if creating a geodemographic classification designed to be used by the public. The similar structure of most commercial classifications means users have become accustomed to such designs, albeit with certain elements unique to each classification. An example of this is the normal practice for classifications to come in a hierarchical

form, with each level of the hierarchy having different number of clusters. The number of levels to the hierarchy follows no strict guidelines, nor do the number of groups in each. Table 2.1 shows the hierarchical structure of the current commercial geodemographic systems available in the UK, along with the number of groups at each level. Classifications such as Acorn and Mosaic both have a three-level hierarchy, but have different numbers of groups at each level, while Cameo has a two-tiered structure consisting of 10 and 57 groups respectively. The general lack of consensus on how to structure or form a geodemographic classification suggests that there is no set rule on the matter. It implies that structure and group numbers are derived more from data availability and user requirements (Singleton and Spielman, 2013) or expectations, rather than any statistical or analytic method.

Academic geodemographic classifications can therefore either replicate these approaches, or instead of re-inventing the wheel, look to the future and attempt something genuinely new. This forward outlook is most keenly seen with research into future potential data sources. The ONS Beyond 2011 programme (ONS, 2013a) has not only got CACI's Acorn contemplating a post-Census era, but academia as well. Recent research into using social media data in geodemographic studies (Adnan et al., 2013) suggests a possible new rich data source. To what extent this may be a passing interest, or a genuinely useful source of data is yet to be determined, but the creative freedom that academia has with no commercial pressures is certainly a benefit in these circumstances. It is however unclear how such data sources could be used in future geodemographic classifications. The trade-off between robust statistics and new data sources, such as social media, that have little or no statistical basis, could lead to the development of new methods for constructing geodemographic classifications.

New geodemographic classifications tend to iterate rather than truly innovate. There is currently a growing movement in both the commercial and academic sides of geodemographics that this legacy view may not be sustainable moving forward. The release of 2011 UK Census data in 2013 and 2014 has checked this view for the time being, as the Census remains the best and most complete source to use in geodemographics, and not using such a valuable resource would be a waste. It would however be inexcusable for any classification created during this time not to have an eye on the future. Even if the latest developments have no direct bearing on the methodologies and data used, it is still important to understand where new classifications will sit within the research field. It would however seem likely the current

round of classifications either in construction or already available, both from the commercial and academic perspectives will be the last ones to have a major reliance on Census data, even if there are future Censuses.

2.6. The 2001 Area Classification for Output Areas

The 2001 Area Classification for Output Areas, also known as the 2001 Output Area Classification (2001 OAC), was created in partnership with the ONS by Daniel Vickers for his PhD. The classification was released in 2005 and has remained unchanged since then. The 2001 OAC is an example of a classification that uses only Census data in its construction, in this case 2001 UK Census data.

The methodology used to create the 2001 OAC was constrained by the requirement for the classification to fit within the wider family of ONS classifications created from the 2001 UK Census, such as those for local authorities, health areas and wards. Initially 94 variables were selected from key statistic tables released by the three Census agencies of the UK; the ONS for England and Wales, the General Register Office for Scotland (GROS), now the National Records of Scotland (NRS), for Scotland and the Northern Ireland Statistics and Research Agency (NISRA) for Northern Ireland. These 94 variables were reduced to 41 variables (see Table 2.2), with Vickers et al. (2005) giving a detailed account of the variable selection process. To summarise, the aim with the original selection process was to decide upon the least number of variables that best represented the main aspects of the UK population as captured by the 2001 UK Census. Variables were chosen to avoid certain characteristics that can have adverse impacts on a geodemographic classification, such as being highly correlated with other variables, have badly behaved distributions, having limited geographic variation and a lack of temporal robustness. The final selection consisted of a combination of individual and composite variables with the limiting long-term illness variable being standardised for the effect of age structure (Vickers et al., 2005).

Table 2.2: Variables used to construct the 2001 OAC

v1	% Age 0 – 4	v22	Rooms per household
v2	% Age 5 -14	v23	People per room
v3	% Age 25 – 44	v24	% HE qualifications
v4	% Age 45 – 64	v25	% Routine/Semi-Routine occupation
v5	% Age 65+	v26	% 2+ Car household
v6	% Indian/Pakistani/Bangladeshi	v27	% Public transport to work
v7	% Black African, Black Caribbean or Black Other	v28	% Work from home
v8	% Born outside UK	v29	% Limiting long term illness rate (Standard Illness Ratio)
v9	Population density	v30	% Provide unpaid care
v10	% Divorced	v31	% Students (full time)
v11	% Single person household (not pensioner)	v32	% Unemployed
v12	% Single pensioner household (pensioner)	v33	% Working part-time
v13	% Lone parent household	v34	% Economically inactive looking after family
v14	% Two adults no children	v35	% Agriculture/fishing employment
v15	% Households with non-dependant children	v36	% Mining/ quarrying/ construction employment
v16	% Rent (public)	v37	% Manufacturing employment
v17	% Rent (private)	v38	% Hotel & catering employment
v18	% Terraced Housing	v39	% Health/social work employment
v19	% Detached Housing	v40	% Financial intermediation employment
v20	% All Flats	v41	% Wholesale/retail employment
v21	% No central heating		

Adapted from Vickers et al. (2005)

After the variable selection process was complete the data had to be prepared before it could be clustered. An outline of the procedure is summarised in Figure 2.5. A decision was made that the classification would not have weighted variables. Vickers and Rees (2007) explain this was a result of wanting the classification to be general purpose, and the unpredictability potential weighting schema would have had on the relationship between variables and the effect this would have on the final classification was deemed to be undesirable. Using untransformed variables led to unsatisfactory clustering solutions so variables were logarithmically transformed after all the values had one added to them to prevent zero values producing errors. This transformation was applied to all variables uniformly, even though for three of the variables their skewness increased. The benefits of transforming the entire dataset outweighed any negative impacts on these individual variables. Variables were then standardised so each variable had the same range and thereby the same weighting, preventing outliers from biasing the clustering process (Vickers et al., 2005). The method utilised was range standardisation, opposed to inter-decile or z-score standardisation. The reasoning given for this choice by Vickers et al. (2005) was that inter-decile and z-score standardisations gave too much weight to highly skewed variables; range standardisation, while itself being susceptible to extreme outliers, was the most effective standardisation method for reducing the impact of outliers (Petersen et al., 2011).

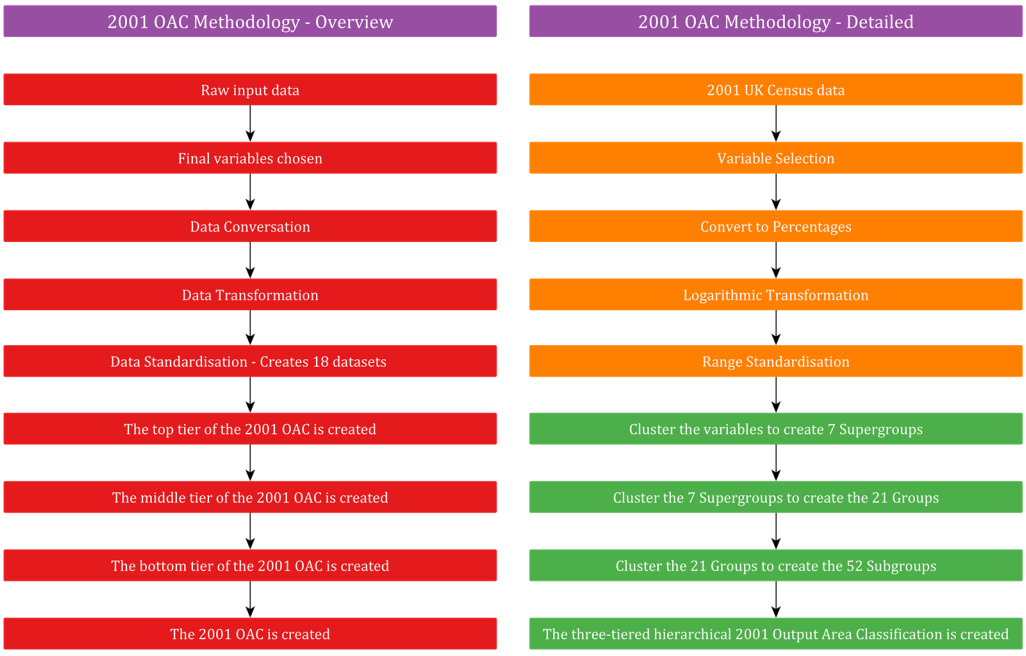


Figure 2.5: Methodology of the 2001 OAC

Adapted from Vickers and Rees (2007)

The 2001 OAC is a three-tiered hierarchical classification. Prior to clustering, the number of groups for each tier were selected. These values (7, 21 and 52) were used as they resulted in the most optimum clusters in terms of composition, geographical distribution and the proportion of Output Areas (OAs) each represented. K-means was the clustering method used, after initial tests using Ward's hierarchical clustering method (Ward, 1963) in conjunction with k-means created non-uniform clusters (Vickers et al., 2005). The use of k-means rather than Ward's hierarchical clustering was significant as it had a direct impact on the structure of the classification. Using Ward's hierarchical clustering would have created a classification from the bottom-up, where each areal unit to be clustered starts off by itself and then gets grouped by lessening degrees of similarity to the point where theoretically all the areas could be grouped together. To create a tiered hierarchy three points in the clustering process could have been chosen that reflected the requirements of the classification. This can be considered a more traditional approach to creating a geodemographic classification.

The alternative method of using k-means to make a hierarchical classification makes the process more computationally intensive and does not naturally lend itself to creating a hierarchy. Using k-means meant creating the top level of the classification, the Supergroup level, first, with 7 groups being found to be the optimum number. Each of these groups was then clustered again individually to form the middle level of the classification, the Group level, consisting of 21 groups. These 21 groups were then clustered individually to form the bottom level, the Subgroup level, consisting of a total of 52 groups. Table 2.3 shows the names assigned to the Supergroups and Groups to summarise the characteristics present within each respective cluster grouping; the Subgroups were assigned subcategorised names derived from the Group name (Vickers et al., 2005).

The naming of any clusters in a geodemographic classification is a potentially challenging task; the names cannot be too ambiguous otherwise they risk serving no useful purpose and blend into each other. They also cannot be too specific, as this would only truly represent a relatively small proportion of the population in each group. With this in mind, the ecological fallacy needs to be a constant consideration to prevent incorrect or false inferences being made of individual-level relationships from area-level data (Robinson, 1950). Figure 2.6 is a choropleth map of the UK showing the Supergroups of the 2001 OAC. For a more detailed explanation of the 2001 OAC methodology see Vickers et al. (2005) and Vickers (2006).

Table 2.3: The 2001 OAC hierarchy

Supergroups	Groups	Subgroups
1 - Blue Collar Communities	1a - Terraced Blue Collar	1a1
		1a2
		1a3
	1b - Young Blue Collar	1b1
		1b2
	1c - Older Blue Collar	1c1
		1c2
		1c3
2 - City Living	2a - Transient Communities	2a1
		2a2
	2b - Settled in the City	2b1
		2b2
3 - Countryside	3a - Village Life	3a1
		3a2
	3b - Agricultural	3b1
		3b2
	3c - Accessible Countryside	3c1
		3c2
4 - Prospering Suburbs	4a - Prospering Younger Families	4a1
		4a2
	4b - Prospering Older Families	4b1
		4b2
		4b3
		4b4
	4c - Prospering Semis	4c1
		4c2
		4c3
	4d - Thriving Suburbs	4d1
		4d2
5 - Constrained by Circumstances	5a - Senior Communities	5a1
		5a2
	5b - Older Workers	5b1
		5b2
		5b3
		5b4
	5c - Public Housing	5c1
		5c2
		5c3
6 - Typical Traits	6a - Settled Households	6a1
		6a2
	6b - Least Divergent	6b1
		6b2
		6b3
	6c - Young Families in Terraced Homes	6c1
		6c2
	6d - Aspiring Households	6d1
		6d2
7 - Multicultural	7a - Asian Communities	7a1
		7a2
		7a3
	7b - African-Caribbean Communities	7b1
		7b2

Adapted from Vickers et al. (2005)

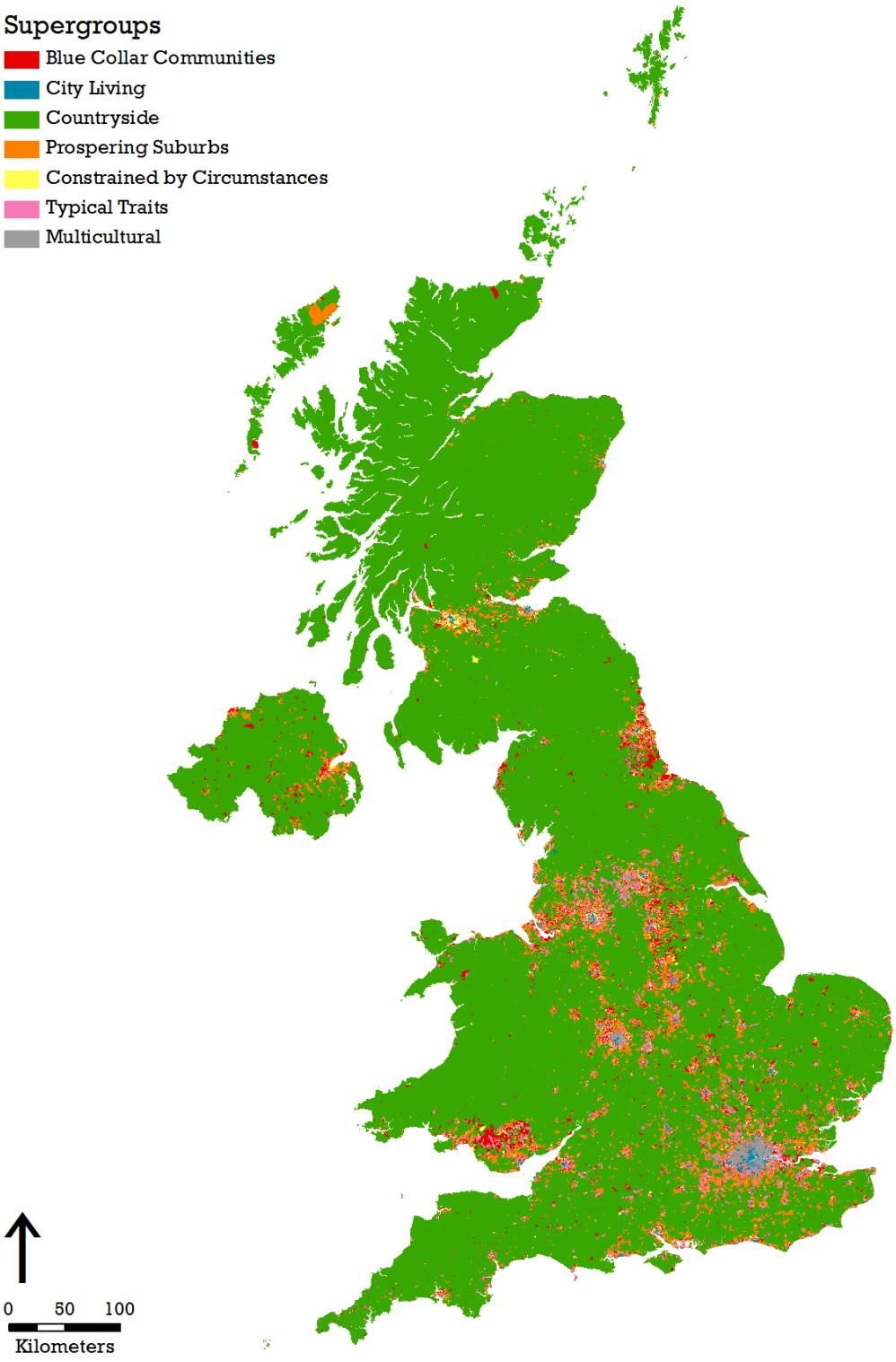


Figure 2.6: The 2001 OAC Supergroups

2.7. London and the 2001 Area Classification for Output Areas

The 2001 OAC has been widely used, helped by the ONS coding some of their statistical outputs with the classification. Local authorities in particular have been keen users of the classification, for example Cambridgeshire County Council coded their Place Survey with the 2001 OAC (Cambridgeshire County Council, 2012). Although the 2001 OAC has been adopted across different geographical areas throughout the UK, there are certain locations that are poorly represented; for example Gale and Longley (2012) identified problems with the 2001 OAC's representation of London.

Table 2.4 shows the assignments of the seven 2001 OAC Supergroups across the UK and in London only. The groups have a more even distribution across the UK compared to London. In London the 'Multicultural' Supergroup represents 56.1% of the population, and the 'City Living' Supergroup another 21.4%. Together these groups represent over a quarter of London's population. Although there are the Groups and Subgroup levels of the 2001 OAC to help drill down the results to provide a more refined representation of neighbourhood conditions, the casual or untrained user seeing over half of London being designated 'Multicultural' may find this vague or even unhelpful. This problem is caused by the high diversity of London being accommodated within a national classification; in important respects the UK is set apart from the prevailing characteristics of its capital city. The numerical size of the 'Multicultural' Supergroup attests to this fact, while the cluster profiles of Vickers et al. (2005) identifies the average numbers of individuals described as 'Indian, Pakistani and Bangladeshi', 'Black' and 'Born Outside UK' as key to defining this Supergroup. In a sense London is forced in with the rest of the UK in the quest to devise a 'one size fits all' classification, a decision that has hindered the effectiveness of the 2001 OAC for anyone wanting to use it in London.

National classifications due to how they are normally constructed are not perfect entities. London however exacerbates the problem due to its special settlement status, with Webber (2007) providing evidence that London is unique within the UK. Webber's study used the functions conceived by Hall et al. (2001) in defining urban centres to rank metropolitan habitus in England; of which London was placed top of multiple hierarchies, in the top strata. Indeed, the assumed importance of London can be identified in the functions used to define the central place rankings of English cities by Hall et al. (2001) as one of them used was "direct train connection to London". While this does not provide any empirical evidence of London's unique status it does show an understanding exists in the literature that London is different to the rest of the UK.

Table 2.4: 2001 OAC Supergroup Distribution

2001 OAC Supergroup	**UK**	**London**
Blue Collar Communities	16.1% (35,837)	2.5% (606)
City Living	7.5% (16,637)	21.4% (5,174)
Countryside	12.4% (27,681)	0.1% (21)
Prospering Suburbs	21.2% (47,250)	7.4% (1,782)
Constrained by Circumstances	14.9% (33,165)	2.5% (592)
Typical Traits	18.3% (40,769)	10.1% (2,430)
Multicultural	9.7% (21,721)	56.1% (13,535)

(Counts are in brackets)

The evidence would suggest therefore that London is a separate entity in geodemographic terms as is demonstrated by Longley et al. (2011). This indicates London would be better suited as being grouped together with other world cities - rather than as a region within the UK. Petersen et al. (2011) suggest that "national classifications tend to be represented regionally by one dominant neighbourhood type" (p. 177) and that the remaining types then diminish exponentially. The steps taken by Vickers et al. (2005) to create the best clustering solution for the 2001 OAC across the UK have meant London has suffered as a consequence in terms of the classification returning useful results for that one particular region. As London does not fit well into a nationwide context this prevents the 2001 OAC from being considered to be a truly representative national classification of the United Kingdom. Petersen et al. (2011) identified two potential methods for dealing with this problem: segment the clusters further or create a regional classification. The solution favoured by Petersen et al. (2011) was to create a regional classification – the 2001 London Output Area Classification (2001 LOAC). This followed the same procedures as those used by Vickers et al. (2005) when creating the 2001 OAC, but used a London only dataset and created new names

and descriptions for the clusters. Conceived at the Supergroup level, the names and frequencies of each group are shown in Table 2.5. The distribution of the cluster assignments of the 2001 LOAC resembles that of 2001 OAC at a national level, as the spread amongst the groups is fairly even. Figures 2.7 and 2.8 are choropleth maps of the 2001 OAC in London and the 2001 LOAC. Visually they are distinct classifications, and at a basic level this can be considered a satisfactory improvement if the desire is for a classification that only represents London and cannot be compared to a national equivalent.

Table 2.5: 2001 LOAC Supergroup Names and Distributions

2001 LOAC Supergroup	**Frequency**	**Percentage**
Suburban	2,506	10.4
Council Flats	3,678	15.2
Asian Quarters	2,716	11.3
Central District	3,409	14.1
Blue Collar	3,114	12.9
City Commuter	3,542	14.7
London Terraces	5,175	21.4
Total	**24,140**	**100.0**

Adapted from Petersen et al. (2011)

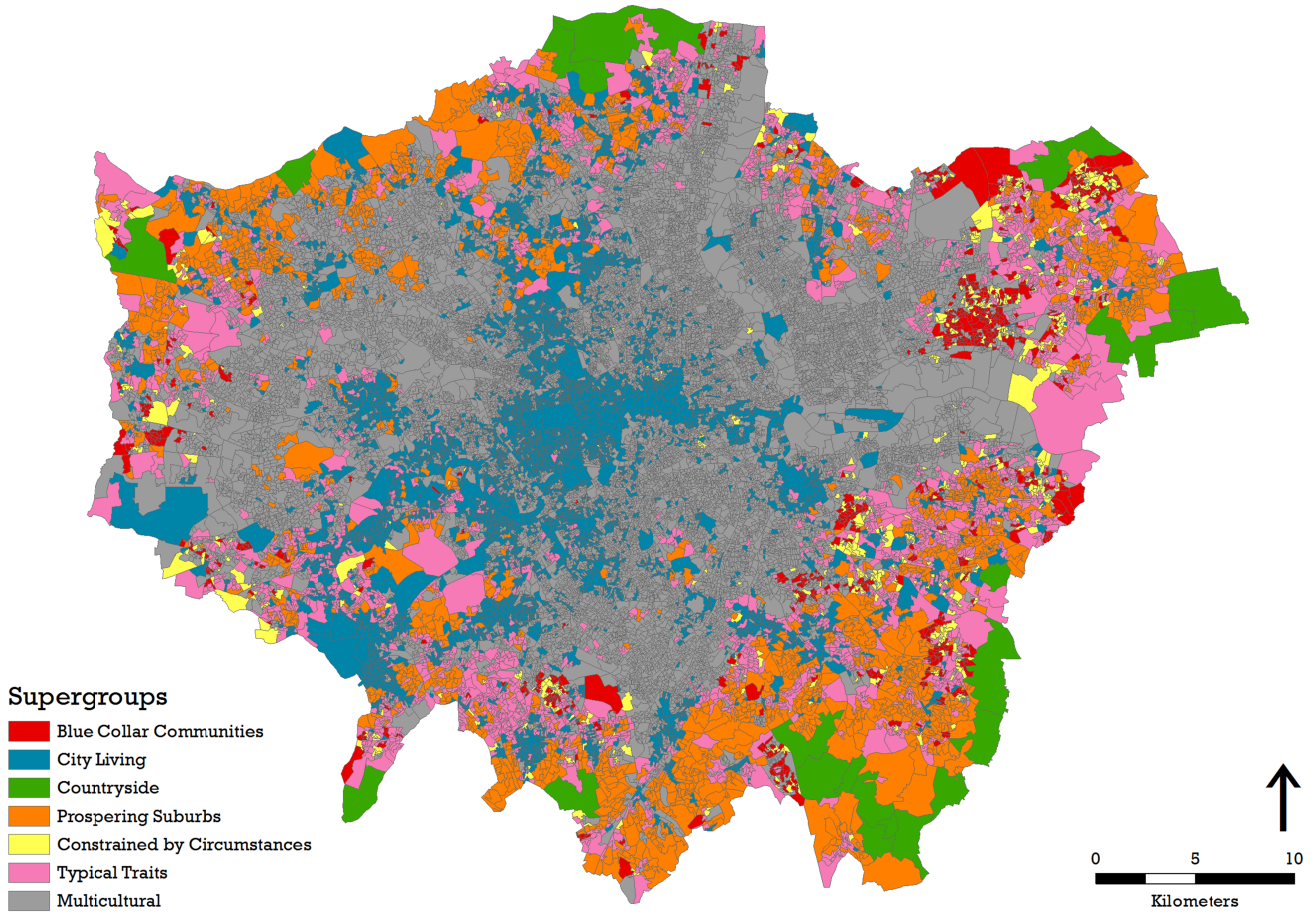


Figure 2.7: The 2001 OAC in London

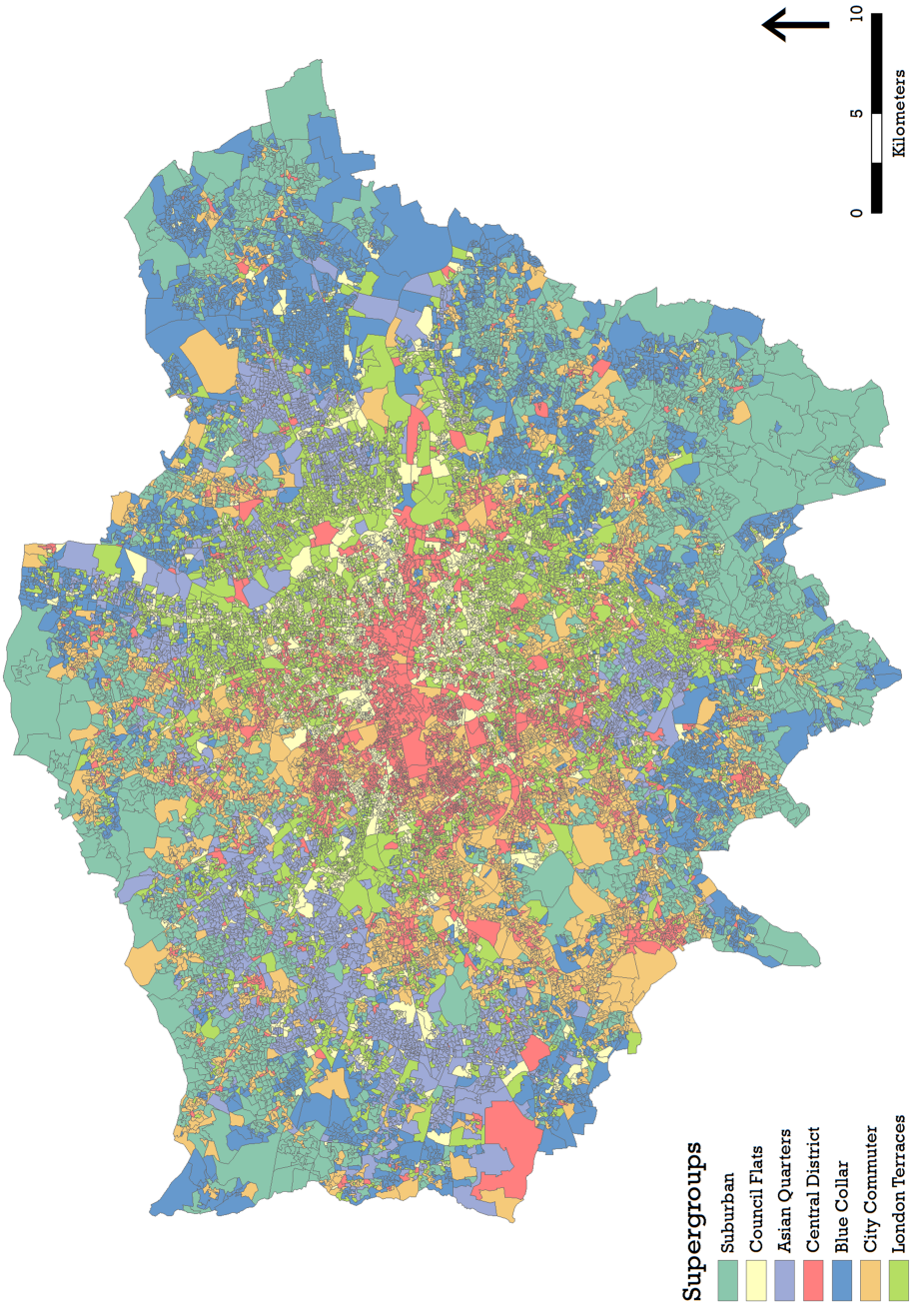


Figure 2.8: The 2001 LOAC

The rationale for the 2001 LOAC provokes wider questions of motivation, specification and estimation of open geodemographics. For example, the motivation to create the 2001 LOAC was based upon the results of the 2001 OAC in London not being as representative as they could have been. The 2001 LOAC is not the only example of a regional geodemographic classification existing for this reason; Kingston upon Hull's city council, as previously mentioned, has created its own classification (Feldman, 2011). In the commercial sector CACI released the research tool 'London Segments' in 2009 (CACI, 2009) that uses the same methodology as its Acorn classification. The developments that have occurred in the time between the release of the 2001 OAC and the 2011 UK Census would indicate the rise in popularity of regional geodemographic classifications. It is likely that this perceived shift is a result of users identifying shortcomings with the national classifications available, and needing something more tailored to their specific needs.

The release of the 2011 UK Census data is likely to see a shift back towards favouring national classifications, at least in the short term. The reasons for creating regional classifications may vary. For example, it can be argued creating the 2001 LOAC was a necessity to have a good geodemographic understanding of London. This may not always be the case, as results from the 2011 UK Census shows that the rest of England and Wales is becoming more ethnically diverse like London (ONS, 2012b). While this would indicate London becoming less unique within England and Wales, it is still likely to exhibit characteristics not found elsewhere in the country for the foreseeable future. The extent to which these characteristics would warrant their own classification remains unknown. How new national classifications cope with London will in part decide if new versions of classifications like the 2001 LOAC are warranted. If they are not needed in the same way the 2001 LOAC was, then these classifications switch from being a necessity to being a complementarity geodemographic tool to national classifications.

2.8. Potential Pitfalls of Geodemographics

Geodemographics has many advantages in the way it summarises multidimensional datasets. These advantages outweigh the negative aspects, as seen by the continued development of geodemographic applications. The negative aspects cannot however be ignored, as they can create barriers in the successful development of geodemographic applications. These issues range from general points that that can apply to any field dealing with spatial data, to more specific points that relate only to geodemographics.

2.8.1. The Ecological Fallacy

Geodemographic classifications are generalisations. The average characteristics derived from multiple inputs are by their nature not individual level data, and this is taken to be the best representation of the resident population. The lack of any information at the individual level means making inferences about a person's characteristics from where they live and not from direct observations. If an individual lives in an area that contains 70% student households this does not mean they themselves are a student, although the chances of this being true are much higher than if they lived in areas that contained only 10%. Additionally, if two variables at the aggregate level had a high correlation, it does not mean the same relationship can be inferred at the individual level. For example if an area has a high level of persons aged over 45 and high levels of marriage, it does not mean every person aged over 45 is married, or that individuals under 45 cannot be married. Furthermore, the same problem can occur when inferring relationships from higher aggregation levels to lower aggregation levels. If a local authority had an unemployment rate of 20%, it does not mean that every town or city that falls within its boundary is the same. The actual rates of unemployment in such locations could in fact be higher or lower than the 20% figure. The only thing known for certain is that the average of all areas within the boundary will equal 20%. The ecological fallacy can therefore be defined as either the erroneous inference of relationships from the aggregate to the individual level, or from any higher aggregation level to a lower aggregation level (Robinson, 1950).

The dangers of the ecological fallacy and geodemographics are well documented in the literature (Birkin, 1995; Harris et al., 2005). By design geodemographics appears guilty of making erroneous assumptions that individuals living in an area share the general characteristics of the overall population. From an ecological fallacy standpoint however it is not geodemographic classifications that cause problems, rather, it is the way they are interpreted. Webber (2007) illustrated that there are distinct groups inhabiting areas of inner London that cannot be found elsewhere in Britain, while Butler and Robson (2003) suggest London's middle classes distinguish themselves from the middle classes of other cities. If such areas do exist at a local scale then any distinctions these residents may make about their own placing on the social hierarchy are unlikely to be represented at a national level. This is why "one should take care to avoid inferring from the geodemographic description that all residents share the overrepresented behavioural characteristics that distinguish one type of neighbourhood from another" (Webber, 2007, p. 206).

Geodemographic classifications operate on the simple fact that areas that have a population with certain characteristics are more likely to contain individuals with the same traits. Deviation away from this interpretation of results is what can create ecological fallacy issues. The only way to avoid any prospect of an ecological fallacy within geodemographics would be to utilise individual, rather than aggregated, data. The difference between using such data sources to form geodemographic classifications is that aggregated data emphasise geographical concentrations, while individual data are aspatial (Openshaw, 1984a) and only available at coarser levels of geography in order to protect respondent confidentiality.

2.8.2. The Modifiable Areal Unit Problem

The Modifiable Areal Unit Problem (MAUP) refers to how raw data can be aggregated in different ways to produce different results. The MAUP phenomenon was first identified by Gehlke and Biehl (1934) who identified that the correlation coefficient of male juvenile delinquency differed depending on the scale of aggregation, however the term MAUP was first coined by Openshaw and Taylor (1979). The literature on the subject has studied the complexities of MAUP (Openshaw, 1984b; Fotheringham and Wong, 1991; Tranmer and Steel, 2001), with it being considered a fundamental geographic problem that can impact all studies of spatially aggregated data (Wrigley, 1995). MAUP itself consists of two components: the scale effect and the zoning or aggregation effect. The scale effect relates to how the scale at which data are represented impacts upon the results of analysis and interpretation and is defined by Jelinski and Wu (1996) as "where the same set of areal data are aggregated into several sets of larger areal units, with each combination leading to different data values and inferences" (p. 129). The zoning effect relates to how a study area is divided, and how different arrangement of zones, even at the same scale with no large variation in size or shape, can lead to different results.

The degree to which the MAUP can impact upon the results of geodemographic classification is dependent on the selected areal units. Openshaw and Rao (1995) achieved correlations between unemployment and households with no car ranging from -1.00 to +1.00 by modifying the size and shape of the areal units. Although, another way of looking at it is that one correct result was hidden by all the incorrect values. The study undertaken by Openshaw and Rao (1995), and many similar, set out to explicitly modify areal units to achieve such seemingly alarming results. It is unlikely that data aggregated

to pre-existing geographies, such as wards, would lead to such extreme results, despite the fact that the majority of these geographies were not optimised for the release of statistical outputs. Geodemographic classifications that utilise Census data released since 2001 in the UK have an advantage in this respect. Prior to 2001, the majority of Census data released at the smallest spatial level used arbitrary spatial units designed for statistical collection, not statistical data distribution. With no natural or meaningful boundaries the size and shape of the spatial units exhibited wide variation in their social homogeneity. As a result, the MAUP would have been a valid concern for any statistical analysis, opening any conclusion drawn to potential criticism. Since 2001 the smallest spatial units used to release UK Census data have been created using an automated zoning procedure (AZP), first developed by Stan Openshaw (1977). This technique creates zones using a set of predetermined requirements (such as minimum population levels), eliminating the ambiguity of arbitrary created equivalents. Using AZP ensures the spatial units used represent the optimum arrangement in terms of size, shape and social homogeneity in the majority of cases.

As a potential source of error that can affect any spatial study that utilises aggregate data sources (Unwin, 1996), MAUP could have a detrimental impact on geodemographic classifications. The use of geographies constructed using AZP help to mitigate any negative impacts in modern day geodemographics. Data that are pre-aggregated to the smallest spatial level of these geographies represents an optimum source. The data cannot be aggregated down to smaller spatial levels, and the final zone design is as optimum as possible (based on the original design criteria). In many aspects utilising these geographies for a geodemographic classification eliminates the negative impacts of MAUP. The creation of standard geographies designed to fulfil a set of clearly defined requirements means there can be no argument they represent an optimal outcome, and therefore can be considered 'correct'. It is however important to note that any changes to the design criteria would likely lead to a completely different set of geographies, which would impact on the composition of a geodemographic classification. MAUP therefore has a greater impact on the creators of geographies, rather than those who use them.

2.8.3. Validity of Geodemographics

The term 'geodemographics' was coined in the 1970's, as discussed in Section 2.3, and over the past 40 years geodemographic classifications have been widely used across the UK. However, some would argue that the research field is not particularly informed by

theory (Flowerdew and Leventhal, 1998) and in the mid-nineties it was recognised that there had been little 'hard evidence' to prove that geodemographics actually 'worked' (Harris et al., 2005). This led the Census and Geodemographics Group (formerly the Census Interest Group) of the Market Research Society to conduct a thorough study, which became known as the Luton Case History and provided evidence to demonstrate geodemographics success. The extent to how much a geodemographic classification may be criticised for lacking scientific theory can be mitigated if the final outputs are perceived to be of use, by providing an accurate representation of a population's characteristics.

The importance of knowing if geodemographics 'works' can be argued to be dependent on who uses the systems. Businesses who use the commercial products do not need evidence to show how a geodemographic classification is formulated or even works, as they only need to justify decisions made internally. Contrast this with academic and public sector users, and justification becomes more important. It is not enough just to have a classification for such users, it is important to know if decisions made can stand up to scrutiny and be backed up by statistical logic.

Some argue a geodemographic classification is a reflection of the past, rather than any current or future conditions. The rationale behind this is that the data used to construct the classifications is always dated before it becomes available. The time between the last two Censuses in the UK being held and the first data release occurring has ranged between 18 months and 2 years, with the full release of all the data occurring around 3 years after Census day. During this period changes caused by the UK's population moving each year are not recorded.

There is a perception that there is a significant negative impact on a classification when it is built using data that could be seen as 'out of date'. This view is emphasised by commercial operators, as they want their regular use of ancillary data sources for updating their classification to be seen as the key advantage of their product. The concept that change is a constant process is correct, but it needs to be defined from a geodemographic perspective. Just because there is large population churn occurring at the smallest spatial level does not automatically equate to a classification becoming invalid. Sleight (2004) suggests this change does not matter as certain types of people will always dominate certain areas and as people move out similar people move in. This accounts for short-term fluctuations in the social compositions of an area. In the longer-

term, the creation of a regional geography of Britain using surnames by Longley et al. (2011) has shown, with the possible exceptions of Britain's urban conurbations, that a large proportion of the British population has remained settled for at least 600 years.

The notion that the denser populated urban areas are likely to experience the most change is unsurprising. The population of London increased by 12% between 2001 and 2011 (ONS, 2012c), but as shown by Orford et al. (2002), temporal change appears to be superficial with little impact on the effectiveness of classifications. Additionally, complete changes to neighbourhood composition and their geo-social hierarchy were unlikely to have happened, except in places where significant redevelopment had occurred resulting in change to the social construction of residents, like that in the London Docklands. Although it seems counterintuitive with so much change occurring, geodemographic classifications can none-the-less be deemed as relatively stable. The relative social patterns of areas are however for the most part stable, and have been stable for decades (Harris et al., 2005). As such, using data that is several years old does not significantly impair the enduring relevance of geodemographic classifications.

2.8.4. User Engagement

Historically, users of geodemographic classifications have had limited say in how such systems were created. The 2001 OAC for example did not include any formal user consultation on how the classification should be formed and what the outputs should be. Recent attempts to rectify this position have been trialled, with the concept of 'consultation shaped geodemographics' found in Longley and Singleton (2009) as part of their e-Society classification. The e-Society classification divides neighbourhoods in Great Britain based on the levels of awareness of information and communications technologies, usage patterns, and attitudes to their effects upon quality of life. Their use of consultation was limited to a validation exercise of the e-Society classification using the Internet. During the 13 day period the consultation ran for the website received 79,051 hits and 3,952 feedback responses, indicating an interest amongst certain members of the public to engage with geodemographics. This achievement should be tempered with the observation that, as Jones (1999) and De Vaus (2002) note, internet surveys are susceptible to self-selection and response bias.

Advertising the existence of geodemographic classifications to the general public creates a different issue, exposing the names and descriptions used for the groups to wider

critique. Vickers (2006) describes this as perhaps the most important part of any classification, and the inevitable consequence of more people having access to classification descriptors is greater disagreement. This highlights the importance of phrasing when deciding upon names and descriptions. The challenge is providing accurate names and descriptions that reflect the likely characteristics of the population, but without potentially causing offence. A geodemographic classification is not to tell a population who they should be, but is a product of who they are, and the names and descriptions should reflect this. The only way this can happen is with public consultation and validation. This process is not repeated by commercial systems, and is an area where academic geodemographics can actively engage with a user base.

A geodemographic classification should be a mixture of consultation on what the classification should be, to make sure you are creating something that is wanted in addition to user validation to ensure the final product meets those requirements. This validation should relate to all aspects such as the number of groups and not just their names and descriptions. Modern geodemographics, especially in academia, are ideally placed to meet these requirements and make the process of creating a classification a much more interactive and cooperative endeavour, with fully documented reasons for decisions made being an important part of an open and transparent agenda. Longley and Singleton (2009) have shown there is a desire for the general public to engage with geodemographics, and Vickers and Rees (2007) have shown that ground-truthing of classifications can play an important role in validating their conclusions. This open and transparent approach, in contrast to the closed approach taken by the commercial operators, demonstrates that just because something works well from a business perspective does not mean alternative approaches to geodemographics cannot exist.

2.9. Conclusions and Research Agenda

Area classification and geodemographic research dates back over 100 years to Charles Booth's study of London towards the end of the 19th century, although it can be argued that the intellectual heritage of geodemographics actually begins in the 1920s and 1930s with the work on urban studies by the Chicago School. These conceptual beginnings of urban ecology and social area analysis provide a framework for social measurement to be empirically undertaken to better understand neighbourhood characteristics. The desire to generalise urban social patterning in the 1970s led Richard Webber to develop a branch of applied urban studies that would later be termed 'geodemographics'. While

initially devised for use in the public sector, successful geodemographic applications developed most rapidly in the private sector, using proprietary solutions such as CACI's Acorn and Experian's Mosaic. Despite their cost, commercial geodemographic products such as these dominate the UK market and are widely utilised across multiple industries, in part because ancillary sources are used to enrich and update the classifications. Many users equate 'frequently updated' with being more accurate and therefore assume these options are the best ones; despite the origins of some of the ancillary sources remaining unknown (Experian, 2010; CACI, 2013a). The provision of incremental updates over time is the focus of much hype within the marketing of commercial classifications. However, the 2001 OAC offers an alternative to the commercial products (Vickers and Rees, 2007). The 2001 OAC is freely available in the public domain providing a national level classification built at the small area level and is the outcome of a well-documented, transparent and easily replicable methodology. The classification relies solely on data from the 2001 UK Census and therefore, unlike the commercial packages, it does not adopt any inter-censal updates..

Vickers and Rees (2007) note that the real advance that the 2001 OAC made was the realisation that such a classification could be built using freely available data and provide a viable alternative to commercial equivalents. Since the release of the 2001 OAC, academic geodemographics has focussed more on bespoke variations of classification in research, some of which has dealt with the problems of the 2001 OAC as identified by Gale and Longley (2012). The release of 2011 UK Census data provides an excellent opportunity to address issues of the 2001 OAC by creating a new national level classification at the smallest spatial level. Consultation and validation are important components of this, and would add more validity to the final system and make the entire process less authoritarian. There is also an opportunity to acknowledge the growing desire to create bespoke variations by creating a methodology that will be freely available, as well as easily adaptable to different datasets and requirements. These concepts also have the benefit of lending themselves to any future developments in geodemographics, such as the use of non-Census Open Data. For concepts such as these to be successful there is a need to move beyond what the 2001 OAC offered in terms of providing a transparent methodology, and actively encourage the practical side of creating geodemographic classifications by using free open source software like R (R Development Core Team, 2011). This would not only allow for bespoke variations of the classification to be made more easily, but should allow reproduction of the new

classification. This forms an important part of academic geodemographics: the ability to critique classifications, both how they are made and their final outputs.

Creating any new geodemographic classification does not mean creating carbon copies of previous systems, or mimicking the features found it current iterations. The use of ancillary data sources to regularly update classifications is the feature that separates commercial and academic geodemographics the most. It should not be the place of academic geodemographics to provide a replica of this approach, and developers should instead seek different ways of dealing with the change that does take place, even if the majority of this is unlikely to have any impact from a geodemographic perspective (Orford et al., 2002; Longley et al., 2011). The creation of a new open geodemographic classification provides an opportunity for methods to be devised to deal with this change. This should act to encourage the continued use of a classification for an extended period after its release, something that did not necessarily happen with the 2001 OAC.

This project is an opportunity to build on the success of the 2001 OAC, address some of the problems identified in the literature, incorporate recent advances in computing and GIS and provide a new geodemographic classification born out of academia that seeks to address the needs of all users and observers of the field and move the research agenda forwards. The main aim of the project can be summarised as creating a new geodemographic classification of the UK at the smallest available spatial units of analysis. This in turn can be further divided into a set of more specific goals:

- Creating a new open, transparent and reproducible methodology that can be adapted and applied in different scenarios to aid the creation of bespoke geodemographic classifications.
- Consult with users to determine what their requirements are for the classification.
- Develop visual and descriptive outputs to facilitate users understanding of the results produced by the classification.
- Validate the classification once complete to assess the final outcome.
- Explore alternatives to using ancillary data sources to update geodemographic classifications that highlight the temporal stability, or otherwise, of resident populations.

Chapter 3

The Census and Open Data

3.1. Introduction

This chapter introduces the UK Census and Open Data as the potential data sources for the new small area classification of the UK. Section 3.2 details the history of the UK Census, and how it is unparalleled as a data source in terms of population coverage and the breadth of questions asked. The different release schedules of the three Census agencies responsible for disseminating the 2011 UK Census are explained, and what this meant for creating the new UK-wide small area classification in a timely manner. The quality of the data is also examined, and the steps taken to guarantee that it is the most robust and accurate dataset available, something which is particularly important from a geodemographic perspective. Section 3.3 looks at the geographies used to disseminate the Census, and how these have changed over time. It explains their current hierarchical nature, and the differences between the three Census agencies. Section 3.4 examines the possibility of the 2011 UK Census being the last of its kind, and what the future may be for small area statistics in the UK, with particular reference to the impact this may have on geodemographic applications.

Section 3.5 explores the future of Open Data in the UK, and how the release of small area statistics form a relatively small part of this politically motivated agenda. Section 3.6 looks at possible data sources that could be available from commercial companies, and how they could be used in geodemographic classifications. Section 3.7 draws together the theme of Open Data and geodemographics, and how it is not as new a concept as it may seem. It also examines how realistic it is to use non-Census Open Data sources at present, and the positive and negative effects this could have on geodemographic classifications. Section 3.8 draws together the points raised about the use of Census data and Open Data, and provides an outline for the data sources used in creating the new small area classification of the UK. Section 3.9 examines the issues around the late release of 2011 Census data for Scotland at the smallest spatial scales, and the impact on creating

the new UK-wide classification. Finally Section 3.10 draws all these points together, indicating how the research can progress.

3.2. The UK Census

The Census is the most complete data source holding demographic and socio-economic characteristics for the UK (UK Data Service, 2013). This information is used by the government, private companies, in academia and by the voluntary sector for multiple purposes. For example: in informing governmental policy-making and funding allocation; in providing a benchmark dataset, in identifying and targeting disadvantaged areas and in informing and supporting research agendas. Information derived from the Census is also a valuable tool for marketing companies and business planners (Raper et al., 1992). This wide use of Census outputs indicates that the Census has continued importance across multiple disciplines, in addition to its value for geodemographics. Almost all UK geodemographic classifications created since the 1970s have utilised some element of Census outputs, and the 2001 OAC relied on the 2001 UK Census as its sole data source (Vickers and Rees, 2007).

The first UK Census was carried out in 1801 and has taken place every ten years since (with the exception of 1941 due to World War II). The most recent Census was conducted on the 27thMarch 2011 by three organisations; the Office for National Statistics (ONS) for England and Wales, National Records of Scotland (NRS – formally the General Register Office for Scotland or GROS), and the Northern Ireland Statistics and Research Agency (NISRA). Although remarkably similar, each Census agency had slight variations in the questions asked. The most notable differences related to the languages spoken, reflecting the geographic variation in the topic across the UK. In England and Wales 56 questions were asked, a 37% increase compared to 2001. There was also an increase in the number of questions asked in Scotland and Northern Ireland, with the addition of questions covering language, national identity and second addresses. Additionally, some questions were expanded, such as ethnic groups, which included 'Arab' for the first time. However, following the 2007 test Census, questions asking respondents directly about their income were dropped (Collins et al., 2010).

The basic set of questions asked led to a complex set of results. The questions had a multitude of possible answers, ranging from only a few options per question to over one hundred. These could be tabulated individually, or cross tabulated with other responses,

creating large multidimensional datasets. Such a vast amount of data requires processing and quality assurance – a time consuming process. Each Census agency releases their data individually, which means that the datasets providing coverage of the entire of the UK are not available simultaneously.

Table 3.1 indicates the release schedule of the Census data across the UK. Release of data from each agency is done in stages, and within these stages are multiple phases. This rolling release schedule allows the simpler forms of data, such as univariate statistics, to be processed and released first; and the more complex datasets, such as the multivariate statistics, to be published later. Variations in this approach do exist between the Census agencies, but each release stage broadly covers the following topics:

- First release: Base population statistics relating to age and sex of residents
- Second release: Univariate statistics
- Third release: Multivariate detailed characteristics
- Fourth release: Multivariate local characteristics

The second release stage was of the most interest to geodemographics applications and to the new small area classification of the UK. Univariate statistics (counts of a single attribute per spatial unit) are the type of data that drive geodemographic classifications. To utilise Census data to create a UK-wide classification requires full geographical coverage at the smallest spatial scale. The first data were released in England and Wales on the 30th January 2013, on the 30th January 2013 and 28th February 2013 in Northern Ireland and the 18th December 2013 in Scotland. This delay in receiving the Scottish data caused a significant delay in the use of the latest Census data to construct the new UK-wide small area classification.

Table 3.1: The release schedule for 2011 UK Census outputs

Country	First Release Stage	Second Release Stage	Third Release Stage	Fourth Release Stage
England and Wales	16th July 2012	11th December 2012	16th May 2013	31st July 2013
Scotland	17th December 2012	26th September 2013	27th February 2014	N/A
Northern Ireland	16th July 2012	11th December 2012	16th May 2013	20th March 2014

(Dates denote first phase of each release)

3.2.1. Data Quality

The quality of Census data is high, but there are several issues that relate to how the data are collected and processed. Firstly, it is a legal requirement to complete the Census in the UK. However, no Census is ever a complete enumeration of the population (Simpson, 2003; Martin, 2010). Knowledge of the issues relating to the overall quality of Census outputs are vital.

The complex nature of social change has made undertaking a modern Census more difficult. As a result, response rates are harder to maintain, with undercounting of the population to be expected (Plewis et al., 2011). The 2001 and 2011 UK Censuses utilised the same basic method of completing a total population enumeration, and then used the Census Coverage Survey (CCS) to adjust for any undercount (ONS, 2012d). The CSS was a small-sample voluntary survey to measure the coverage of each Census. It was used to estimate the population counted and missed, and the final counts were adjusted accordingly (ONS, 2012e). The CCS acts to reduce bias population estimates, as those who were originally missed by each Census are likely to be different from those who were enumerated (Plewis et al., 2011). In addition to this, imputation was required for responses that were incomplete.

The 2001 and 2011 UK Censuses shared the same basic methodology, but differed procedurally in areas such as data collection. This is in part due to inadequacies identified with the enumeration of the 2001 UK Census, with the most noticeable problems occurring in Manchester and the London Borough of Westminster. These two locations suffered large under-enumeration, with 26,200 and 17,500 missing people for Manchester and Westminster respectively (ONS, 2004a). Reasons for this significant under-enumeration in Westminster include difficulty in finding and accessing properties, a low awareness of the Census by the population, and a significant proportion of the population not speaking English as their first language (Pharoah and Rowe, 2010). Issues with 2001 UK Census enumeration were not isolated to these two locations. Trends such as young people, especially young males, being less likely to complete their forms when compared to other parts of society were identified (Simpson, 2002). Additionally, individuals who were in the UK illegally were less likely to complete their forms due to fear that the data would be used against them (ONS, 2005). The overall response rates were 94% in England and Wales; however, 12 Local Authorities had response rates lower than 80%. These Local Authorities were all boroughs in London, with Kensington and Chelsea having the lowest response rate at 64% (ONS, 2004b),

suggesting that urban areas are more likely to see patterns of low enumeration in comparison to rural locations.

The 2011 UK Census aimed to achieve better coverage than the previous Census at the Local Authority level, by addressing previous enumeration failures. The overall target in England and Wales was to achieve a 94% response rate, with at least an 80% response rate for each Local Authority (ONS, 2012f). To achieve this the ONS moved away from enumerators hand-delivering Census forms, to using a central address register and the Royal Mail to post forms. Additionally, the Census form could be completed online for the first time; although only 16% of responses were submitted this way (ONS, 2012g). These less labour intensive methods (in terms of enumerator fieldwork) utilised for the 2011 UK Census, coupled with knowledge that 92% of Local Authorities achieved response rates of 90% or higher for the 2001 Census, meant the ONS could allocate increased resources to areas that were predefined as the hardest to enumerate in England and Wales. Areas, such as Westminster, therefore, had tailored approaches that took into consideration the issues likely to cause under-enumeration. This targeting of problem areas was effective as the response rates for Local Authorities in England and Wales ranged from 82% to 98%, with an overall response rate of 94%. However, these headline figures can hide large variations within Local Authorities themselves, with some areas likely to have had much lower overall response rates than 80% (Plewis et al., 2011). Responses also varied by the categories described in ONS (2012f), with the lowest response rate coming from residents in Basildon, Essex who lived in a caravan or other mobile or temporary structure, where only 16% responded.

High response rates are key to accurate estimation of the number and location of those missed by the Census (Abbott, 2009). The proactive approach used in the 2011 UK Census to obtain universally higher response rates meant that extreme cases of under-enumeration as seen in Manchester and Westminster in 2001 were not repeated, and as a result less imputation was required. The increased reliance on true respondent data, rather than estimated data can therefore be considered to create a more accurate dataset from the 2011 UK Census in comparison to previous Censuses.

Over counting is another issue that can impact the quality of the Census data. This can occur when an individual is counted as being resident at multiple addresses, or if a fictitious person is recorded as residing in a household on Census day. Historically, overcount has not been considered a significant issue in England and Wales (ONS,

2012h). Results from the 2001 Census in England and Wales suggest a double counting rate of 0.4%, or 1 in 250 residents, with two-thirds of these likely to be caused by students having been enumerated as resident at their home and term-time addresses (Plewis et al., 2011). It was however anticipated that the 2011 UK Census would show increased instances of overcount, with estimations ranging from 0.5% to 1% (300,000 to 600,000 individuals), with the greatest propensity again caused by students (Large and Brown, 2010). The basis of these assumptions were largely due to changes in data collection methods and social behaviour, such as second homes causing duplications (ONS, 2012h). The final estimation of overcount in England and Wales was 0.6%, or 352,000 people (ONS, 2012h), considerably smaller than the 6% of missing respondents, but a figure which still required adjustment in the final estimations.

The overall quality of Census data is therefore a product of both under enumeration and over counting of residents. Following the 2011 UK Census the net under coverage for England and Wales was around 5.4% before any adjustments were made as a result of the CSS or any other quality assurance process. The increased response rates, and relatively low increase in over counting means, in England and Wales at least, that the 2011 UK Census can be considered more robust and accurate when compared to previous Censuses. As a data source detailing the general characteristics of the population, it cannot be surpassed and is ideal for continued use in geodemographics.

3.3. UK Census Geography

As the outputs of any Census are inherently geographic, spatial units that adequately reflect this are required to disseminate results. Disseminating the data of any Census is a complex task with a myriad of geographies available to use, each with their own unique hierarchical compositions. Hierarchies such as administrative, electoral and postal all fulfil different purposes, and vary in degrees of suitability for disseminating statistical content. Electoral geography for example is designed to facilitate polling in neighbourhoods at the finest granularity, with different constituencies used for local, national and EU elections. The boundaries of such constituencies are often controversial (The Boundary Commission for England, 2013). Spatial units designed to facilitate effective hand delivery of Census questionnaires by Census enumerators, ‘enumeration districts’, were also used to report Census results up until 1991 in England, Wales and Northern Ireland, and 1981 in Scotland. Enumeration districts (EDs) were thus used for both data collection and publishing the results of each Census. EDs were designed to nest

within the higher administrative geographies of wards (postcode sectors in Scotland) and parishes (communities in Wales) (Martin, 2002a), and to equalise the workloads of the enumerators covering each area. A result of this was that EDs, like many other areal units, bore only a limited relationship to the social, economic and demographic distributions of characteristics of local populations (Openshaw, 1984a), and exhibited large variations in size and social homogeneity (Martin, 2000). This meant that each ED would often span marked social divisions (Morphet, 1993).

The unsatisfactory nature of EDs led to the creation of new spatial units for the 2001 UK Census, called Output Areas (OAs). Unlike EDs they were created after the Census took place and were designed specifically for the statistical release of data. They formed the smallest spatial element of Census geography for the 2001 UK Census, and therefore became the primary unit of dissemination. A geographical hierarchy was formed (see Figure 3.1) by the aggregation of OAs together to create Super Output Areas (SOAs). In England and Wales these were split into two layers, Lower Layer Super Output Areas (LSOAs) and Middle Layer Super Output Areas (MSOAs). A third level, Upper Layer Super Output Levels (USOAs) were also made available for Wales. In Northern Ireland solely SOAs were constructed. In Scotland a similar approach of aggregating OAs together was undertaken to produce Data Zones (DZs) and Intermediate Zones. These geographies have remained for the 2011 UK Census, with only minor alterations and changes in terminology taking place.

An understanding of Census data geographies is important when considering them as a source for geodemographic classifications, as data at the finest spatial scale are preferable. UK Census geography remains complex, mainly due to the three different agencies responsible for creating their own variants. Some differences exist (see Sections 3.3.1 and 3.3.2), and knowledge of what these are, both between countries and within the hierarchies is important (Rees et al., 2002). It is also important to consider non-Census datasets that are made available using the same Census geographies. While Census geographies were born out of the Census, they are now used as the default spatial units for disseminating other data as well, made available by the government and statistical agencies of the UK through websites like Neighbourhood Statistics. An important consideration is therefore how these multiple datasets that use Census geographies coexist, and how they can be utilised most effectively.

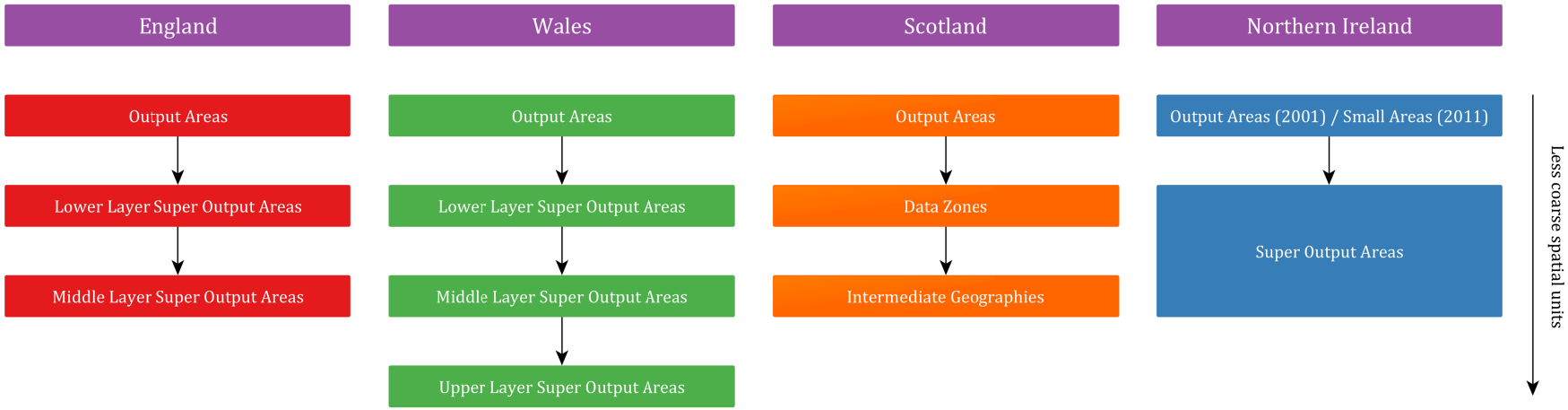


Figure 3.1: UK Census Geography in 2001 and 2011

3.3.1. Output Areas and Small Areas

Output Areas (OAs) were first created by the GROS for use with the 1991 Census in Scotland. They were built from postcodes, and designed to match the EDs used in previous Censuses as closely as possible. Built using a GIS and manual intervention, they were considered for use in England and Wales, but the Office for Population, Censuses and Surveys (OPCS) concluded that it would be ineffectual to reproduce such a manual and time consuming process for a much larger set of postcodes (Vickers, 2006).

The solution for England and Wales was the adaption of the automated zoning procedure (AZP) (Openshaw, 1977) by Professor David Martin. Using an AZP method, adjacent postcode areas in England and Wales could be grouped together, depending on set criteria, to provide full and continuous geographical coverage. Three statistical criteria were used to optimise the OAs: 1) population size controls to reduce inter-OA variance, 2) maximise social homogeneity, and 3) the shape must be as compact and as circular as possible. In addition to these criteria, OAs were designed to be constrained by obvious boundaries, such as major roads, and to nest within administrative geographies. Thresholds were also placed on the minimum numbers of residents and households per OA in an effort to ensure confidentiality of the data. The AZP used the set criteria and threshold values to swap postcode areas between OAs until an optimum was reached. For a full overview of the methodology used to create OAs, see Martin (1998, 2000, 2002a, 2002b) and Martin et al. (2001).

The 2001 UK Census provided the first opportunity for OAs to be used by all three Census agencies. Although they were designed and constructed after the Census took place unlike the previous EDs (Martin, 2002a), their use in dissemination across the UK was not uniform. This was due to the three Census agencies being responsible for their own zoning procedures. As such, there were differences between the agencies (ONS, 2013b), with the ONS and the NISRA adopting a different minimum number of residents and households to the GROS (see Table 3.2). The differences found in Scotland were a result of a desire to keep the 2001 OAs as comparable with the 1991 OAs as possible. In addition to this, OAs in Scotland were explicitly designed to avoid mixing urban and rural locations together. The result of these methodological differences can be seen in Table 3.2, where the variations in the average population and number of households can be seen between the constituent countries of the UK.

Table 3.2: 2001 Output Areas

Country	2001 OAs	Population	Households	Average Population per 2001 OA	Average Households per 2001 OA
UK	223,060	58,789,194	24,479,439	264	110
England and Wales	175,434	52,041,916	21,660,475	297 (100 minimum)	124 (40 minimum)
Scotland	42,604	5,062,011	2,192,246	119 (50 minimum)	52 (20 minimum)
Northern Ireland	5,022	1,685,267	626,718	336 (100 minimum)	125 (40 minimum)

Source: 2001 UK Census

The use of the AZP to create OAs was no doubt an improvement for dissemination purposes compared to the previously used EDs, but did create some issues. OAs provide full geographical coverage, but the Census only relates to the population, and more specifically the night-time locations of the population. As such, non-residential areas, such as industrial parks, had no population data, but were still assigned to an OA. In order to meet the set criteria and meet the minimum thresholds, this resulted in very large (in terms of geographical coverage) OAs being formed. This was also true for any geographical features that had no resident population, such as large water bodies, as they still needed to be attached to a minimum number of residents and households. The opposite problem existed for large populations that shared the same postcode unit, such as university halls of residence, where because of the shared postcode they could not be broken down into smaller OAs. Finally, because OAs apportion a 3D world into 2D shapes, this caused particular problems with tower blocks which typically house large populations in a small geographic area that cannot be divided accurately in two dimensions. Each tower block was therefore given its own OA, irrespective of social homogeneity, leaving certain OAs with very high population densities and likely candidates for outliers in geodemographic classifications (Martin, 2002b). These issues often meant that the geographically larger 2001 OAs tended to have smaller populations than the geographically smaller ones; the counterintuitive nature of this can therefore be problematic when it comes to visualising any data that uses them.

The problems with the 2001 OAs, while important, should not detract from their successful deployment. The use of a spatial unit designed for dissemination purposes is pivotal in reducing the role of the Modifiable Areal Unit Problem (MAUP) – see Section 2.8.2 – in research that followed the release of data from the 2001 UK Census. They also provided a foundation on which these spatial units could be updated for use in disseminating the 2011 UK Census data. Updating, rather than constructing entirely new spatial units, provided an opportunity for consistency in the geographies used to report Censuses for the first time. However, it is important to note again that differences remain between countries due to the individual agencies responsible for the data. At the start of 2013 the 2011 OAs were released for England and Wales and Northern Ireland, although in Northern Ireland were renamed Small Areas (SAs). The 2011 OAs for Scotland were released in December 2013.

The majority of the 2011 OAs in England and Wales are identical to their 2001 counterparts; however there are some exceptions where significant changes, such as large alterations to the population size, have taken place in the intervening decade. For further information on why some OAs required modifying, see ONS (2012i). The ONS set a target to keep changes below 5% of the total number of 2001 OAs, and depending on the nature of the change 2001 OAs were split or merged with neighbouring OAs. Splits were considered when a 2001 OA had over 650 residents or 50 households. In some circumstances splitting an OA was not possible, so populations have remained above the 650 size threshold. Merges were used when a 2001 OA fell below 100 residents or 40 households, a detailed methodology on how merges and splits were applied can be found in Cockings et al. (2011). These adjustments led to 2.6% of OAs changing from 2001 to 2011, with 1.8% being split and 0.6% being merged. This resulted in a 3.4% increase in the total number of OAs covering England and Wales, rising from 175,434 in 2001 to 181,408 in 2011. The average population and number of households have also risen from 297 and 124 respectively in 2001, to 309 and 129 in 2011.

The 2011 OAs in Scotland are broadly similar to their 2001 and 1991 predecessors. The main aim was continuity where possible, leading the majority of the OAs being the same size between 2001 and 2011. Merges were used when a 2001 OA fell below 50 residents or 20 households, and splits when the number of households reached a maximum threshold of approximately 78 households (GROS, 2013). In total there has been an increase of 8.8% OAs from 42,604 in 2001, to 46,351 in 2011. This has resulted in the

average number of usual residents per OA falling from 119 in 2001 to 114 in 2011, although the number of households has remained the same at 51 (NRS, 2013a).

SAs are the new terminology for the smallest spatial level of Census geography in Northern Ireland. They are modified versions of their 2001 OA predecessors and there has been a decrease from 5,022 OAs in 2001, to 4,537 SAs in 2011 – equating to a 9.66% reduction. This decrease is the result of OAs being merged due to population sizes being too small to meet the threshold values, boundary changes, and inaccuracies being rectified (NISRA, 2013). In total 4,175 SAs are the same as their older OA counterparts, representing a consistency rate of 83.13%. However, because of the small number of SAs, the average population and number of households has risen from 336 and 125 respectively in 2001, to 400 and 155 in 2011.

The compatibility between data released at 2001 OA and 2011 OA or SA level is important for two reasons. Firstly, it allows for almost direct comparison between the 2001 UK Census and the 2011 UK Census. This, while obviously having many benefits, is not relevant to the construction of any new small area classification of the UK that does not incorporate elements of change over this time period. The second benefit is more relevant. Compatibility between the two geographies means non-Census data released at OA level prior to 2013 can be incorporated with the newer Census data that utilised the latest OAs or SAs. Data released at the 2001 OA level, such as mid-year population estimates (MYEs) or Council Tax data, can be amalgamated with data released at the 2011 OA or SA level. From a geodemographic perspective this provides more options when determining which data sources can be used.

3.3.2. Other UK Census Geographies

The use of OAs and SAs to release Census data is appropriate due to the format of the Census. Data are collected at the individual and household level, and then aggregated up to the finest spatial level available that does not break data disclosure controls. This is not the case for every non-Census data source, where data collected usually only offer partial coverage of the population and require some element of generalisation. As a result data cannot be accurately disseminated at the most granular spatial scales, either because of inherent uncertainties with the aggregation of data down to these levels or because of data disclosure controls. This therefore requires alternative geographies to OAs and SAs to be used.

Prior to 2001 small area statistics were released at the ward level (the equivalent in Scotland was postcode sectors). These caused problems due to ward populations ranging from fewer than 100 residents to over 30,000, with boundaries also liable to change. This meant comparisons, both at a national level and over time were not possible. Additionally, areas with the smallest populations could not be used as the data would fail disclosure controls.

The solution to using inconsistent and unstable geographies was the introduction of new spatial units following the 2001 UK Census. These higher levels of UK Census geography were built from OAs, and like OAs, each of the three Census agencies differed in the approaches taken to create spatial units with a consistent number of residents and households. In England and Wales, and Northern Ireland SOAs were created.

In England and Wales these were split into LSOAs and MSOAS. LSOAs were typically built from four to six OAs, while MSOAs were larger and contained 25 OAs on average and fitted within local authority boundaries. The thresholds used to construct them varied between 1,000 and 3,000 residents for LSOAs to between 5,000 and 15,000 for MSOAs, while for households it was 400 to 1,200 for LSOAs and 2,000 to 6,000 for MSOAs.

In Northern Ireland the SOAs were not split and had populations that ranged from 1,300 to 2,800, and were constrained to existing wards where possible. The GROS created their own equivalent to SOAs, with their LSOAs being called DZs, and their MSOAs being called Intermediate Zones. The Scottish geographies differed in size compared to their English and Welsh counterparts, with DZs having a population range of 500 to 1,000, while the Intermediate Zones populations ranged from 2,500 to 6,000.

Consistency of these spatial units, like OAs, is another key element of the newer higher-level Census geographies. The same methodology used with the 2011 OAs of splitting and merging existing spatial units was employed if the population and household thresholds were no longer being met. In England and Wales there was a 1.1% increase in the total number of LSOAs, and a 0.1% increase in the total number of MSOAs. In Scotland, a 6.7% increase in the total number of DZ's is proposed, with the number of Intermediate Zones remaining constant. The final changes will be confirmed when these updated geographies are made publicly available in October 2014. In Northern Ireland there were also no changes to the number of SOAs, although some did have their boundaries modified. As Table 3.3 shows however, the small or no change in number of

units can translate to relatively large changes in the mean populations. In England and Wales the increases in residents mirrors the increase in households. In Scotland there is consistency between the mean population and number of households between 2001 and 2011, however in Northern Ireland there has been a large increase in both, especially in the number of households compared to the overall population. The relative consistency of statistically based areal units provides an important base for the release of official statistics. SOAs have become the base spatial unit for the release for the majority of National Statistics, although the outputs of the 2001 and 2011 Censuses have also been released at this level.

Table 3.3: 2011 Super Output Area Geography Populations

	2011 LSOA Population	**2011 LSOA Households**	**2011 MSOA Population**	**2011 MSOA Households**
England and Wales	1,614 (6.61%)	672 (6.66%)	7,787 (7.63%)	3,245 (7.77%)
Scotland*	763 (-0.02%)	342 (0.01%)	4,288 (0.05%)	1,921 (0.08%)
Northern Ireland**	2,035 (7.74%)	778 (11.94%)	Not Applicable	Not Applicable

(Percentage change from 2001 in brackets)

* The LSOA equivalent is DZs and the MSOA equivalent is Intermediate Zones. Values are based on draft proposals for the number of 2011 DZs and Intermediate Zones.

** The LSOA equivalent is SOAs and there is no equivalent to MSOAs.

3.4. Beyond 2011

The 2011 UK Census may have been the last of its kind. In 2008 the Treasury Select Committee published the report 'Counting the Population' (House of Commons Treasury Committee, 2008), making the recommendation that:

> *"the Statistics Authority set strategic objectives to ensure that the data gathered throughout the UK can be used to produce annual population statistics that are of a quality that will enable the 2011 Census to be the last Census in the UK where the population is counted through the collection of Census forms." (para. 149)*

Implementing this recommendation would not be a simple task. The Census has been used as the base for many population and socio-demographic statistics for decades. At present there is no suitable alternative that can provide comparable information at national and local levels on a range of topics and at the finest spatial levels. It has however been recognised that undertaking a Census is becoming increasingly challenging. The current economic climate can make the £480 million cost of the 2011 Census in England and Wales appear expensive and difficult to justify (although, over a decade this equates to costing each individual only 85p a year). Changes in society, with an increasingly mobile population, also mean the concept of producing a snapshot every ten years is becoming less relevant (Benton et al., 2013).

In response to the Treasury Select Committee's recommendation, the ONS initiated the Beyond 2011 programme in April 2011, with NRS following suit in September 2011 (NRS, 2013b), and NISRA also undertaking a similar initiative. The three reviews are being undertaken independently, although there is cooperation between the agencies, hopefully leading to an element of harmonisation in any future UK statistics that do not involve traditional Censuses.

The principle objective of Beyond 2011 is to investigate the best ways of producing population and small area socio-demographic information that best meets the needs of users (NRS, 2013b; ONS, 2013c). This wide brief allows investigation into the many possible solutions to the problem, including retention of the current Census. It also acknowledges the wealth of information that is already being collected by government departments. These administrative records often have full coverage of their relevant populations (Dugmore et al., 2011) and use the same data collection methods across

their geographic extent. Table 3.4 outlines some of the data already being collected by a selection of government departments; while this list is not comprehensive it does indicate the breadth of data that is already held. You cannot however solely utilise these sources as a direct Census replacement. The geographic extent of the records will vary with no ability to link records across departments. It is obviously helpful that such rich data sources exists, but the Beyond 2011 programme is seeking to identify how such sources can be utilised to achieve the wider goals set.

As of July 2013 the Beyond 2011 programme in England and Wales had shortlisted six alternative approaches to the traditional Census as detailed in ONS (2013c):

1. A Full Census. Taken every ten years but with more emphasis on Internet collection.
2. Rolling Census. An annual Census of up to a tenth of the population carried out in different areas each year. This would be similar to the approach currently used in France.
3. Short Form Census and 4% Annual Survey. A Census with a reduced number of questions undertaken every ten years. Supplemented by an annual rolling survey of 4% of the population. Similar to the approach currently used in the USA.
4. Annual Linkage and 10% 10-yearly Survey. Linking administrative data and supplementing it with a decennial survey of a tenth of the population.
5. Annual Linkage and 4% Annual Survey. Linking administrative data and supplementing it with an annual rolling survey of 4% of the population.
6. Annual Linkage and 40% 10-yearly Survey. Linking administrative data and supplementing it with a decennial survey of 40% of the population.

These six options have been designed to reduce field expenditure by utilising online completion where possible, but would differ in the quality of the outputs and total costs. Options 2, 3, 4 and 6 have been discounted as they would result in the loss of the most detailed population statistics, and not lead to any significant cost savings (ONS, 2013c). This leaves Options 1 and 5 as the two approaches the ONS will investigate in detail before a final report is submitted to the UK government in 2014. At this point a decision will be made about what future the Census has in England and Wales, Scotland and Northern Ireland.

Table 3.4: Information collected by government departments

Government department	Database	Possible topics
National Health Service	National Health Service Central Register: GP Patient Registers	• Date of birth • Sex • Address and changes • Births • Marriages • Deaths • Health condition
Department of Work and Pensions; HM Revenue & Customs	Customer Information System (this includes children at birth, as well as the adult population)	• Date of birth • Sex • Marital status • Name of employer • National Insurance (working population and workers from overseas) • Income • Benefits (various, including child allowance, retirement pensions, disability) • Household structure
Department for Education	Annual School Census	• Date of birth • Sex • Language • Ethnicity • Free school meals • Travel to school • Educational attainment
Home Office	e-Borders	• Passport details • Citizenship • International migration
Driver and Vehicle Licensing Agency	Driving Licence	• Address and changes • Car ownership
TV Licensing	TV Licences	• Address • Households
Ministry of Justice/Registry Trust	County Court Judgments	• Personal debt
Valuation Office Agency	Council Tax Bands for domestic properties	• Property value
ONS	Inter Departmental Business Register	• Workplaces and working populations

Adapted from Dugmore et al. (2011)

It is encouraging from a geodemographic perspective that the ONS is pursuing one option that will maintain the status quo of small area statistics. A modernised decennial Census supplemented by administrative data to take account of population change, will still provide similar outputs that the 2011 UK Census provided, but only every decade. Utilising existing administrative data to estimate the population, supplemented with a 4% rolling annual survey to provide estimations of the populations characteristics, would provide data more frequently, but would not have the same level of granular detail as either the 2011 UK Census, or the modernised Census option (ONS, 2013c). At present the precise level of detail possible is not known, but it likely that this option will result in the elimination of most small area data outputs.

The Beyond 2011 programmes provide an insight into the future, or lack of, for small area statistics in the UK. It is clear that there is a desire to retain some element of this, but the extent to which this will be possible will be determined by the desire for spatially granular data, rather than temporally granular data, from the user community before the government makes a final decision. It is clear that the 2011 UK Census was the last of its kind, and this will have an impact on future geodemographic applications. It is difficult to judge in the period between the 2011 UK Census and any future incarnation to what extent Census outputs can be relied upon, and how much time should be invested using alternative data sources. It is likely there will eventually become a tipping point when Census data are not the default dataset for geodemographic classifications. This point will be brought forward if the future Censuses are not able to release data at the smallest spatial levels. If this were the case then it would be the end for traditional geodemographic classifications and alternative solutions would be required. Exploration into the current and future spatial availability of data sources is therefore required, with particular attention on how the release of administrative data fits into the wider context of government policy regarding free and Open Data, and the impact this could have on future geodemographic classifications.

3.5. Open Data

The UK Government, along with other national and regional governments, has been pursuing an Open Data agenda since 2009 (Deloitte, 2012), motivated in large part by the desire to improve transparency in various aspects of government decision-making. There are two elements to the Open Data agenda; firstly making data accessible, and secondly making it useable. In this context, making data accessible refers to publishing

resources that aid understanding in how the government functions and how policies are made. To this end, in January 2010, the website data.gov.uk was launched, acting as a portal where publically available data could be accessed. Ensuring data are useable relates to making public data available in machine-readable formats, published using open standards and released under an open licence (Sheridan and Tennison, 2010). In September 2010 the UK Government launched the 'Open Government Licence' allowing public sector data released under it to be used without need for payment or permission. This license was updated to version 2.0 in June 2013 (The National Archives, 2013).

The Open Data agenda is much broader than simply aiming to release the population and socio-economic data detailed in Table 3.4. It is more closely aligned with a transparent government agenda, designed to correspond with the six opportunities of Open Data identified as: accountability, choice, productivity, quality and outcomes, social growth and economic growth (HM Government, 2011). The release of geographic data, and more specifically spatially referenced data, forms only a small part of the overall scheme at present. The majority of the 9,900 datasets currently available from data.gov.uk (in August 2013) have no spatial reference, and those that do tend to be at higher levels of granularity. It is unlikely that data will ever be released at the individual or postcode unit level due to data disclosure controls, but more datasets could be released at the OA level. At present, aside from Census outputs, there are relatively few data sources released at OA level. The data that are available at OA level are:

- Mid-Year Population Estimates. Released by the ONS annually for England and Wales and by the NISRA for Northern Ireland. The latest release for Northern Ireland is currently for 2008.
- Dwelling Stock by Council Tax Band. Released by the Valuation Office Agency (VOA) for England and Wales annually.
- Workless benefit claimants. Released by the Department for Work and Pensions (DWP) for England and Wales quarterly.
- Land use statistics (Generalised Land Use Database). Released by the Department for Communities and Local Government for England. The last release was for 2005.
- Reported crime data. Released through the data.police.uk website monthly. Available at street level for England, Wales and Northern Ireland and can be aggregated to OA or SA level.

While these sources are limited, with the exception of the MYEs, they are easily accessible. MYEs are available via request from the ONS as the figures are considered experimental statistics and potentially unreliable. The NRS only make them available at the DZ level, while the NISRA do make the data available to download at the OA level, but only have data between 2001 and 2008 currently accessible. This inconsistency at UK level is symptomatic of the current ad-hoc approach to Open Data. The data sources detailed in Table 3.4 could be made available at coarser levels of granularity if there were some form of cohesive strategy regarding the release of spatial data. A recent report, The Shakespeare Review, looked at the future of opening up data for the benefit of government, business and citizens. It concluded that Open Data has the potential to contribute £2 billion to the UK economy in the short term, and up to £7 billion in the future (Shakespeare, 2013). To achieve this, a cohesive approach is required, which is currently missing from most releases of Open Data in the UK, especially in the context of spatial data. An important point made in The Shakespeare Review regarding the quality of data seems particularly pertinent to the UK's statistical bodies; it is suggested that data should be released quickly and imperfectly, and core datasets should subsequently aim to achieve higher quality at a later point. This would seem to contradict the underlying philosophy of organisations like the ONS, who, as seen from their stance on MYEs, are reluctant to release potentially imperfect data.

Making data available, even if they are imperfect, is key to the successful uptake of Open Data and adding value to them. The Ordnance Survey (OS), the mapping agency of Great Britain, is an example of an organisation that has now released multiple datasets. Using the OS OpenData License, a variation of the Open Government License, a user can now utilise certain OS resources any way they see fit for free with no restrictions. This has obvious advantages for users, but also benefits the economy of Great Britain as a whole. It is estimated that the OS OpenData initiative will create a £13 million to £28.5 million increase in Great Britain's gross domestic product (GDP) by 2016 (Carpenter and Watts, 2013), with much of this attributed to the fact that the data are free. Practical uses of Open Data can be seen from the free release of public transport timetables and up-to-date travel information by Transport for London (TfL). Notably, this has led to the creation of a market for mobile phone travel applications. These products, a mixture of free and paid applications, could not exist in the pre-Open Data era. The wide scale release of Open Data in particular fields can therefore be seen as leading a renaissance in how that data are used.

However, when comparing these successes in the Open Data movement with the wider current availability of small area statistics, it is clear work is needed. The OS and TfL examples suggest the release of more spatial data at the smallest spatial level, will, by this act alone, enable new and innovative uses, with geodemographic applications being only one aspect. The fact this does not exist at present reinforces the view that the continued undertaking of a Census is needed. Releasing more data provides an opportunity to shift this standpoint. Making multiple datasets available at the smallest spatial levels ultimately should lead to less reliance on the Census as a data source.

3.6. Commercial Open Data

The data that are held by government departments are undoubtedly extensive. These are not, however, the only sources that hold information about the population of the UK. Commercial companies that provide services to the general public are likely to have large databases of their customers (Dugmore et al., 2011). Table 3.5 gives a brief overview of different commercial sectors and the information they are likely to hold. The statistical robustness of these data sources will vary, although in any case they are unlikely to be completely representative of a resident population. This therefore means individually they are not ideal to be utilised for National Statistics. The records that private companies hold are specific to their market. The level to which they are updated is also a reflection of the types of services offered. What may be lacking in spatial detail, can normally be contrasted with the number of records held. With many companies holding over 10 million records (Dugmore, 2009; Dugmore et al., 2011) providing a detailed view of customer behaviours and interactions.

These records can therefore be considered a potential source for deriving population characteristics, in some cases, down to the individual level. This seems particularly useful from a geodemographics perspective, but less so for agencies like the ONS for whom consistent spatial coverage is a priority.

Table 3.5: Population data collected by commercial companies

Sector	Data they are likely to hold
Retail	• Records of sales to the public of a huge range of products • Sales channels: superstores and local shops, but also online, catalogues, etc. • Major companies often have 10–15 million customers • Limited demographics collected at time of application • Loyalty cards track spending in great detail
Financial Services	• Wide range of products, e.g. current account, mortgage, savings, loans • Various sales channels – branches, ATMs, online, post, etc. • Several companies have over 10 million customers • Detailed demographics collected for some products, for example mortgages • Current accounts and credit cards track spending in great detail • Pooling of databases is well established, e.g. mortgages, savings, credit, fraud
Electricity (and Gas)	• Large coverage (electricity 100% and gas 80%) • Company coverage across the UK is often regional • Minimal demographic information • Much effort is put into maintaining address/meter files • Good data on fraud and debt • Pooling of databases is well established – meter list used by ONS for 2011 Census to identify multi-occupied addresses; Department of Energy and Climate Change statistics on energy consumption
Water	• Each water company has its own territory • Many properties are still billed according to rateable value rather than metered • A great deal of effort is put into maintaining address files • Minimal demographic information • Good data on debt
Telecoms	• Mobile telephone and broadband now has three main players, each with over 15 million customers • Mobiles – Post Pay (monthly contract – an application form is filled in) • Mobiles – Pre Pay (little information collected) • Address information – only basic postal address for 50% of customers on contract • Transaction information has full detail of every call, including location

Adapted from Dugmore (2009) and Dugmore et al. (2011)

The extent to which these commercial sources can be considered 'Open Data' depends on the definition used. The Shakespeare Review suggests that Open Data need not mean 'free' or 'open to everyone', indicating that if companies charged for access to their data then these sources could still be considered 'Open Data'. This contradicts the definition from opendefinition.org, "a piece of data or content is open if anyone, is free to use, reuse, and redistribute it – subject only, at most, to the requirement to attribute and/or share-alike". This is the definition used by Deloitte (a consultancy firm whose research underpins The Shakespeare Review). This viewpoint means that the majority of the data held by commercial companies cannot currently be considered 'Open Data'.

It is difficult to envisage a situation where data held by commercial companies will be made available for free. The examples given in section 3.5 regarding the use of Open Data to boost the national economy does not translate to individual companies that have large data stores. If they were to be released, any commercial value derived by each company would be eroded or lost. This highlights a fundamental difference in attitudes to Open Data. It is becoming apparent that data are more valuable to both users and governments if they are freely available. It is equally true that data are more valuable to commercial companies when they are either locked away or made available for sale. This ideological divide is unlikely to be resolved anytime soon.

3.7. Open Data and Geodemographics

The discussions of an Open Data agenda and the creation of websites like data.gov.uk give the impression that the concept is relatively new. The 2001 OAC is an example of a geodemographic classification that used only Open Data, as the 2001 Census was the first in the UK to be made freely available without a £250,000 license. Other than the Census, Open Data sources have often been discounted from use in academic geodemographic applications, such as the 2001 OAC (Vickers et al., 2005). This is primarily due to the varying spatial coverage and scale at which such sources are available. It is however clear that such inconsistences are not sustainable in the long term. The 'Beyond 2011' programmes gives an indication that future Open Data sources are going to come increasingly from a mixture of administrative sources and possibly from commercial companies, although the extent these could be considered 'open' is debatable. At present however, there are only a relatively small number of spatially referenced datasets, especially at the smallest spatial levels, as discussed in Section 3.5. This means that the

data currently available are better suited to supplementing datasets like the 2011 UK Census, rather than forming the basis of a classification built at the smallest spatial level.

The potential use of alternative Open Data sources to the Census in academic geodemographics is determined by two factors; the desire to use data that either measure something that the Census does not, or allows for the update of classifications. This presents an interesting decision when determining what data should be used to create a geodemographic classification. The Census is the most comprehensive dataset available at the smallest spatial scale. As such, using Census data at an OA and SA level would appear the best choice for a geodemographic classification. Problems do however exist with the limiting decennial release and the lack of supplementary data sources currently available at an OA or SA level. At higher spatial levels there are more data sources available, with a mixture of Census, National Statistics and other Open Data datasets. This greater quantity of sources means a classification can be built that covers topics not included in the Census and that could be regularly updated. The extent to which this offsets the loss of analysis at the finest spatial levels is dependent on the requirements of the geodemographic classification. The choice in current academic geodemographics is therefore between classifications built at the most granular spatial scale, which have the finest resolution but are unlikely to be updated as the Census is the primary data source; or classifications built at higher spatial scales that lack the fine granularity but can include variables not included in the Census and are easier to update during their lifetime.

The problem with using additional Open Data sources, explicitly because they cover a topic not covered by the Census or to update a geodemographic classification, is one of data quality. The problems with such data sources that prevented them being used in the 2001 OAC have not changed. The spatial granularity of Open Data sources is variable, and not all conform to standard Census geographies. The lack of consistency means either the data are aggregated up to a higher spatial level, such as MSOAs, or is modelled down to a smaller spatial level. Neither of these approaches is ideal, as aggregating the data up reduces the resolution of the final classification, and modelling down introduces too many uncertainties for the results to be reliable. A combination of Open Data sources using Census and non-Census geographies is similarly not ideal: Vickers (2003) states that no system exists to transfer data between overlapping or disjointed spatial units to a satisfactory level of accuracy. There is also the issue of spatial coverage. The variations found in some Open Data sources mean that the provision of full UK-wide coverage for

any topic can be difficult to achieve. With different national governments, government departments and statistical bodies responsible for the dissemination of data, releases are often sporadic and sometimes non-existent for certain parts of the UK at the required spatial level. This is not a desirable characteristic when a national classification is required, but becomes less of an issue if only regional or local variations are needed.

The issue of data integrity can also impact Open Data that is available from non-government department sources, such as local authorities or commercial companies. At present, National Statistics and other data from government departments are likely to have gone thorough data quality assurance checks. This means utilising such sources in geodemographic applications provides a certain level of confidence that the final results will provide an accurate representation. Using Open Data from alternative sources does not necessarily bring these same assurances, either in how the data are sampled or how they are generalised to provide wider coverage. Using such data in geodemographics therefore introduces a level of uncertainty not currently found when utilising Census data. With inconsistencies varying between each Open Data source, it becomes a time consuming operation to evaluate the appropriateness of each source for any desired task.

3.8. Data sources for the new classification

It is clear that non-Census Open Data sources will grow in prominence in the future, especially if ONS's Beyond 2011 programme leads to a reliance on the linking of administrative data. Beyond 2011 also provides an indication that the 2011 UK Census is perhaps the last opportunity to utilise such a robust dataset for geodemographic applications. The 2011 UK Census is the most complete Census ever undertaken in the UK, both in terms of total coverage and accuracy, which no combination of alternative Open Data sources currently available can replicate. As such, it is the most suitable dataset for the new small area classification of the UK. The issues associated with using Census data that is several years old by the time it is released were discussed in Section 2.8.3. Although this is not ideal, the impact from a geodemographic perspective is minimal. Utilising non-Census Open Data sources instead to create the new small area classification of the UK would be possible, but it would need to be created at the SOA level, rather than the OA or SA level, due the greater number of datasets being available at this level of granularity. There would however, be underlying issues with the quality of data as discussed in Section 3.7 that may make such as option unworkable in the

present data environment. The other option of supplementing Census data with Open Data sources at the OA or SA level is possible; however the classification would suffer from the lack of full-UK coverage. Creating the new small area classification of the UK from Census data can therefore be considered beneficial as it remains the most comprehensive dataset available to create a geodemographic system. This would however mean the new small area classification of the UK could not be updated, and this may be an issue for potential users.

The ideal academic geodemographic classification would: provide full national coverage; utilise a statistically robust data source (i.e. the Census); offer the finest resolution possible; be updated on a regular basis and use additional variables that are not included as part of the Census. However, this is not achievable within the current UK data landscape. Consequently, it is important to know what qualities are valued most highly by users. The ONS and UCL therefore held a joint user engagement on the new small area classification of the UK. This in part sought to identify the data sources the classification should use, and what other qualities of academic geodemographics were valued by potential users. The results of this are reported in Chapter 4, along with the overall guidance provided for creating the new classification. This user engagement took place before the extent of the delays to the release of the 2011 Census data in Scotland was fully known. In order to move the project forward it was important to know how to deal with the delayed release of 2011 Census data for Scotland at the OA level.

3.9. The new classification and Scotland

The delays to the release of the Scottish 2011 Census data, discussed in Section 3.2, created a problem in building a UK-wide small area geodemographic classification in a timely manner. No combination of Open Data sources could have been used as a direct replacement for Census data, or act as a proxy for it. Alternative possibilities of mixing of 2011 Census data for England, Wales and Northern Ireland with 2001 Census data for Scotland provided a potential solution. However, as the range of Census questions expanded in 2011 (see Section 3.2), any new questions would have had to have been eliminated to guarantee consistency between the datasets. This was felt to be an unsatisfactory solution. A potential alternative to this was to utilise the 2001 Census data and model responses to the new questions. This was considered, but it was felt that the uncertainties this would introduce would more than offset the benefits of producing a UK-wide classification a few months earlier. It was therefore decided not to add any

Scottish data into the new small area classification until the Scottish Census data were released in December 2013.

The data at OA level needed for England and Wales were released in January 2013, and at SA level for Northern Ireland in January and February 2013. This staggered data release of data for Northern Ireland meant that data processing of the England and Wales data had been on-going for one month, and as such it was decided to exclude Northern Ireland from the initial data enquires. As Northern Ireland only constituted 2.25% of the UK's 2001 OAs, compared to 78.65% for England and Wales, it was felt that evaluating the Northern Ireland data as well would require a lot of work, with very limited benefits due to the small increase in coverage. Using only England and Wales data provided an opportunity to evaluate the methodology detailed in Chapter 6 that was used on the new UK-wide small area classification.

The validity of this initial evaluation was however dependant on understanding the relative importance Scotland can have on the final outcomes of a geodemographic classification. This analysis was performed with the 2001 OAC, comparing the full UK version with variations that remove geographic regions to provide a reliable indicator of how much impact regions such as Scotland and Northern Ireland have on a national classification. To achieve this, the 2001 OAC methodology was applied to datasets missing England, Scotland, Wales and Northern Ireland respectively, with the resulting Supergroups analysed both in terms of distribution and composition. Table 3.6 compares the differences between assignments of the 2001 OAC Supergroups. It shows that there are only small changes in the assignments to the Supergroups when any country is removed. However, the bigger the area (in terms of number of OAs), the more susceptible to change it is, but it is only a 2001 OAC with England removed that shows any significant change in Supergroup distribution. It was therefore unlikely that the distribution or number of Supergroups would change with the latter addition of Scotland and Northern Ireland data.

Table 3.6: Change in 2001 OAC Supergroup distributions with geographic regions removed

2001 OAC Supergroup	United Kingdom	Without England	Without Scotland	Without Wales	Without Northern Ireland
Blue Collar Communities	16.1%	+6.47%	-2.63%	-0.06%	-0.14%
City Living	7.5%	-0.15%	+0.93%	-0.02%	+0.05%
Countryside	12.4%	-0.39%	+1.08%	-0.12%	-0.07%
Prospering Suburbs	21.2%	-2.49%	+0.53%	-0.07%	+0.03%
Constrained by Circumstances	14.9%	-4.33%	+0.23%	-0.09%	+0.06%
Typical Traits	18.3%	-2.61%	-0.91%	+0.04%	-0.12%
Multicultural	9.7%	+3.49%	-0.78%	+0.33%	+0.19%

The distribution of the Supergroups is only one aspect; there is also their composition to consider. The larger the number of OAs removed, the bigger the impact is likely to be on how the Supergroups form. Figures 3.2 to 3.5 show how removing countries from the 2001 OAC dataset and re-clustering the remaining data impacts upon the composition of the clusters when compared to the full UK-wide classification. The bars represent the compositions of each Supergroup, and how far above or below the national average each variable is (for a list of variables see Table 2.2). The red part of the bars represents the 2001 OAC, and the blue the 2001 OAC without a certain country. The amount of red or blue depicted by each Supergroup indicates how the composition differs between the two datasets, with bars that are mainly purple suggesting the composition has experienced limited, if any, change.

Figure 3.2, comparing the 2001 OAC with and without England, indicates the largest differences. The 'Multicultural' Supergroup in particular without an England dataset lacks the main drivers of the group's characteristics, namely above average ethnic variables. The change is so pronounced that the group would need a new name and descriptions to best describe its characteristics, although the other six Supergroups all show differences, albeit in smaller amounts. The results shown in Figure 3.2 when compared to Figures 3.3, 3.4 and 3.5 indicate that removing data for Scotland, Wales and Northern Ireland does not have any significant impact of the compositions of the 2001 OAC Supergroups. Figure 3.5 for Northern Ireland shows that the Supergroups are almost identical, and Figure 3.3 for Scotland shows only relatively small changes for Supergroups like 'City Living' and 'Constrained by Circumstances'. These changes are not so significant to change the general composition of the groups, meaning that the names and descriptions given are still likely to be valid.

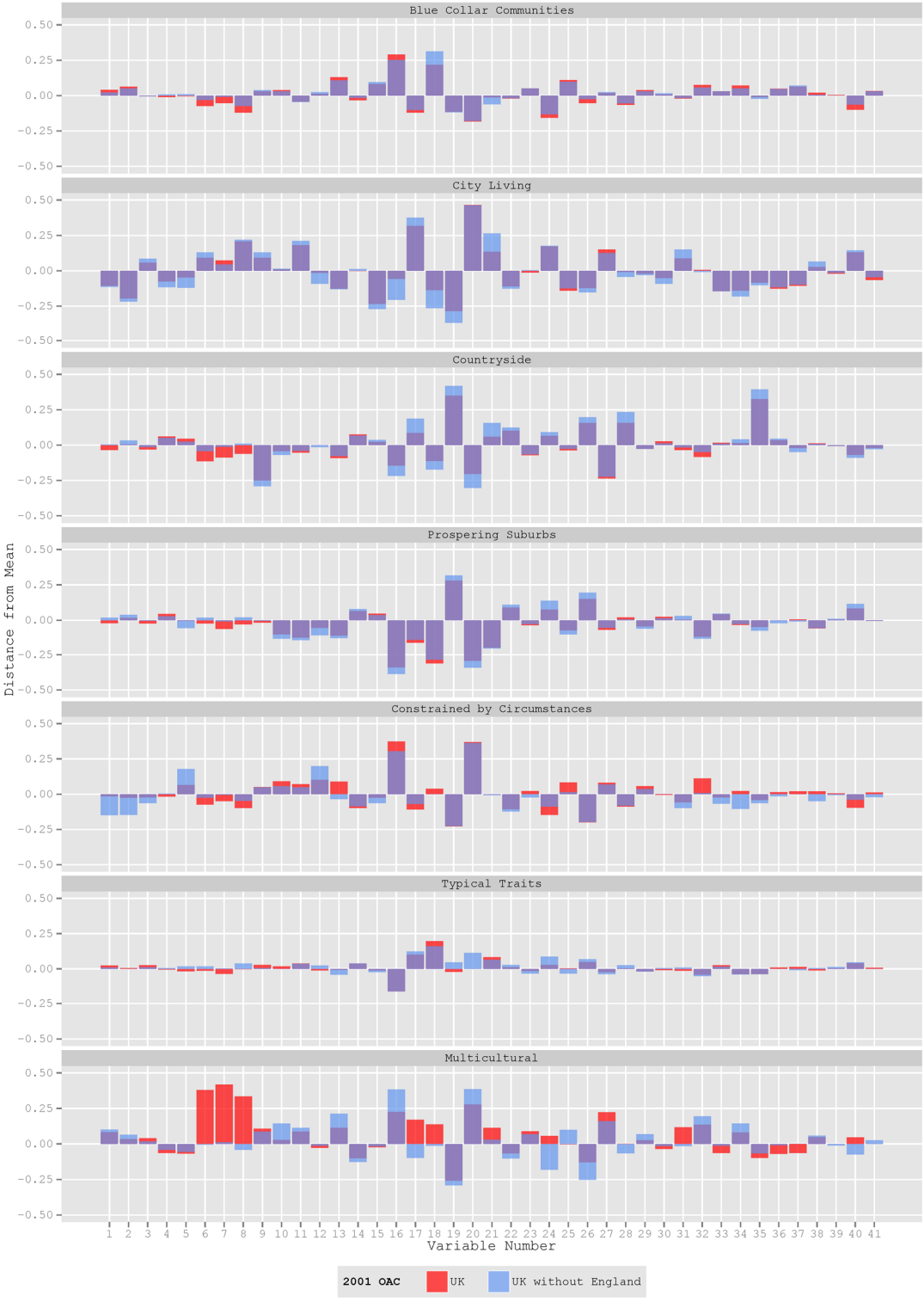


Figure 3.2: The 2001 OAC Supergroup Compositions with and without England

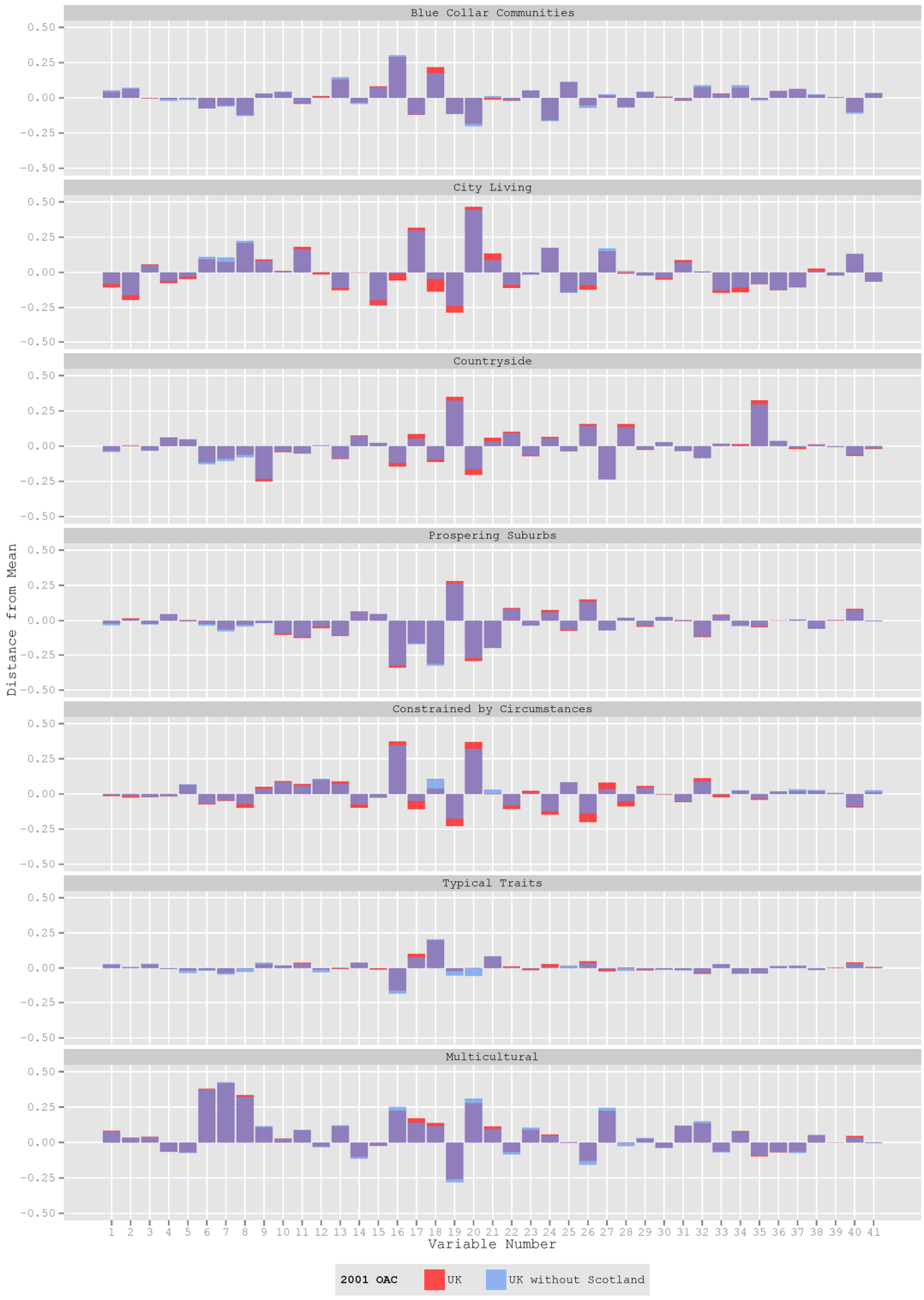


Figure 3.3: The 2001 OAC Supergroup Compositions with and without Scotland

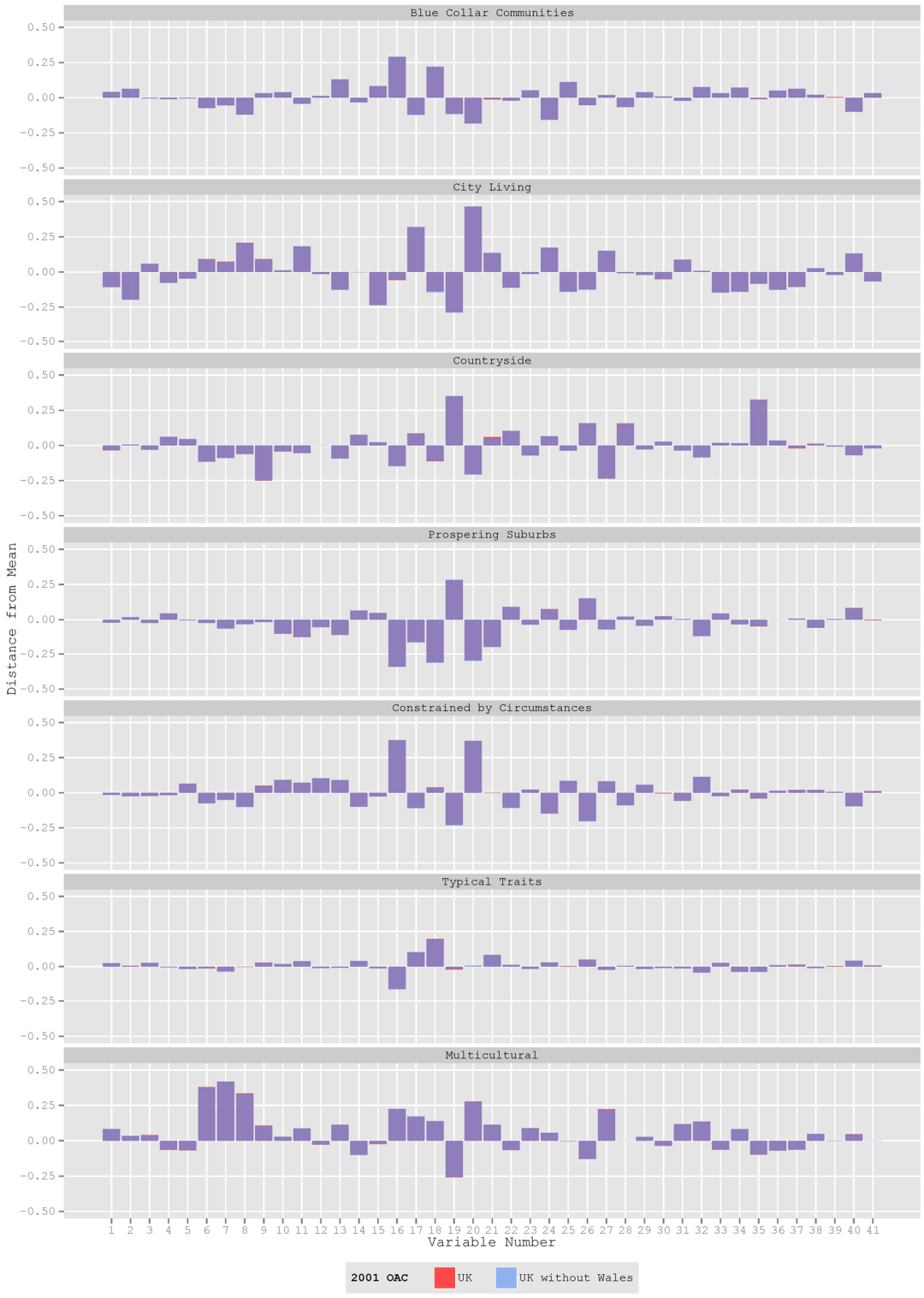


Figure 3.4: The 2001 OAC Supergroup Compositions with and without Wales

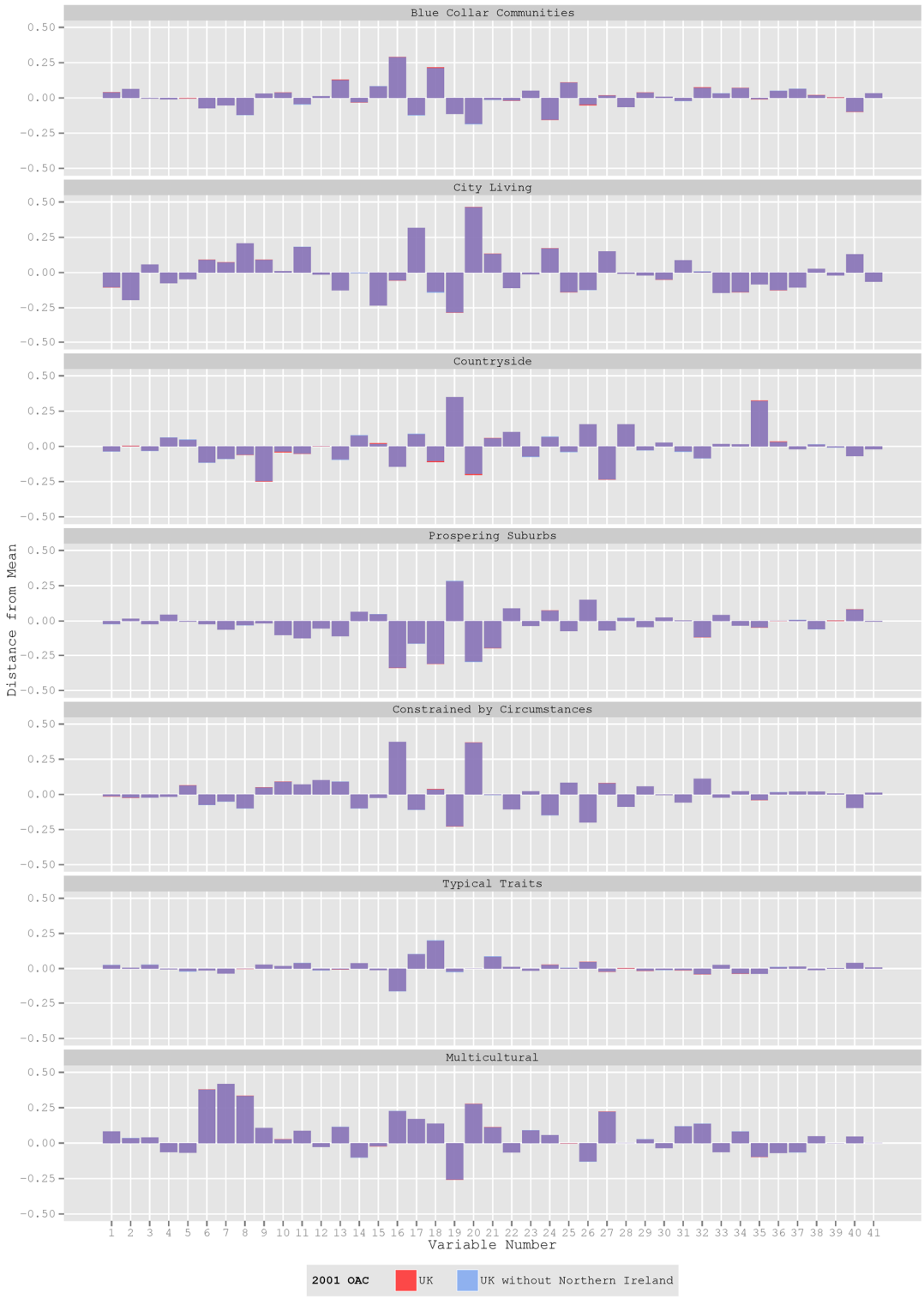


Figure 3.5: The 2001 OAC Supergroup Compositions with and without Northern Ireland

The unsurprising conclusion to draw from this is that England is dominant in shaping the 2001 OAC at a UK national level – it does, after all, comprise 74.27% of the UK's 2001 OAs, and as such accounts for much of the observed heterogeneity in UK population characteristics. The implications for the new UK-wide small area classification were clear. As long as an England dataset was included, any conclusions drawn from evaluating the methodology could be applied when the remainder of the UK data was incorporated. Although this assumption was based on research undertaken with only 2001 OAC data at the Supergroup level, meaning the groups were at their most generic and less susceptible to change. If the research had been also repeated at the Group or Subgroup level then the group compositions were likely to have been more volatile when removing geographic regions due to the smaller proportion of the population they represent.

3.10. Conclusions

The last UK Census, in 2011, was the 22nd since its inception in 1801, and continues to be the most comprehensive geodemographic data source available in the UK, both in terms of coverage and the variety of questions asked. It has formed the basis of almost all geodemographic classifications created in the UK since the 1970s, with the 2001 OAC being an example of a recent classification that relies on it solely as a data source. A reason why it is so widely used, is the quality of the data. The 2001 and 2011 Census in England and Wales had responses from 94% of the population, although this figure masks differences at Local Authority level between the two Censuses. In 2001 twelve Local Authorities, all in London, had response rates lower than 80%. This is problematic, as high response rates are important to accurate Census estimations. In addition to this, the areas of Manchester and Westminster suffered from large under-enumeration due to enumerators having issues gaining access to households. To combat these problems the 2011 UK Census was the first to post Census forms and allow completion online, allowing enumerators to concentrate their efforts on areas identified as hard to count. This practice was deemed successful as response rates for Local Authorities in England and Wales ranged from 82% to 98%, improving the overall accuracy of the 2011 UK Census when compared to those previously undertaken. Therefore the Census can still be considered the default source for geodemographic classifications, especially those built in academia.

The Beyond 2011 programmes are looking to evaluate options if the 2011 UK Census is indeed to be the last. It is likely that there will be no more traditional Censuses, and it will be replaced by linking more administrative data and either a decennial online orientated Census, or by a rolling annual survey that covers 4% of the population. The linking of administrative data fits into the government's Open Data agenda, with the possibility of more spatially referenced data being made available in the future. At present there is a lack of such data at the smallest spatial level, OAs (or SAs for Northern Ireland). OAs and now SAs have formed the basis of Census geography in the UK since 2001, and unlike their predecessors (EDs), were specifically designed for disseminating data and not data collection. Combining OAs or SA together forms higher spatial levels of Census geography, SOAs and DZs, which, in addition to also having Census results, are the basis on which the majority of National Statistics are released. The outcome of this is that more Open Data sources are currently available at higher spatial levels, providing a conundrum when building geodemographic classifications. There is a choice of either constructing a classification at the smallest spatial level with the finest resolution using a data source that cannot be updated; or building at a higher spatial level with data that can potentially be updated and covering additional topics to the Census, but loses the finer level of resolution.

The new small area classification of the UK therefore needed to reflect one of two conflicting approaches. As the 2011 UK Census is likely to be the last of its kind, it would seem a waste not to use it to its full potential. The issue of not being able to update the classification is one that may matter to potential users of the classification, so while a preference should exist for using Census data at the smallest spatial level to build the new classification, the final choice needs to be a reflection of user requirements. Creating any form of new UK-wide small area classification in a timely manner however was hindered by the delayed release of 2011 Census data for Scotland. Possible options were considered on how to proceed, but it was felt waiting for the release of the data was the best course of action. Using the 2001 OAC, it was shown that Scotland, Wales and Northern Ireland are not the main drivers of the classification, with England being the most important component. This meant construction of the new small area classification could be evaluated using data for only England and Wales, in the knowledge that Scotland and Northern Ireland could be added later with minimum impact on the optimum number of clusters and their compositions.

Chapter 4

A New Area Classification for the UK

4.1. Introduction

This chapter introduces the 2011 Area Classification for Output Areas (2011 OAC) and outlines the key concepts that influenced its construction. Section 4.2 defines what the key concepts are and how they are driven both by internal factors, such as the desire for the classification to be open, and external factors like the availability of data. Section 4.3 explains why engaging with potential users was an important component of the 2011 OAC, and what methods were considered to ensure the 2011 OAC user engagement was successfully accomplished. Section 4.3.2 reports the findings of the 2011 OAC user engagement and how the findings influence the general design criteria for the new classification. Section 4.4 summarises the key points from the 2011 OAC user engagement and derives six user requirements from it. These six user requirements are explored further along with the steps taken to incorporate them into the 2011 OAC. Finally, Section 4.5 draws the key concepts and user requirements together. The combination of these two complementary factors is examined, as is their influence upon the design, construction, implementation, use and perception of the 2011 OAC.

4.2. Key Concepts for a new Area Classification for Output Areas

The new small area classification of the UK was named the '2011 Area Classification for Output Areas' (or 2011 OAC) by the Office for National Statistics (ONS). The 2011 OAC shares the basic fundamental principles that govern most past and current geodemographic classifications. There are however key concepts that distinguish the 2011 OAC from the range of other classifications offered in the commercial sector and academia. These concepts are driven both by internal and external factors. Internal factors include the philosophy of the classification, such as a desire for it to have an open and transparent agenda. This forms the fundamental basis of the project, while external factors relate more to the practical elements of the classification. Factors such as the structure and data used can be considered external, and are the more flexible

components of the 2011 OAC. It is however important that these external factors incorporate the views and requirements of potential users. Understanding what would-be users of the 2011 OAC require before its construction is fundamental for increasing uptake once it is released. To help achieve this the views of the wider public were obtained via a user engagement (see Section 4.3).

The key concepts of the 2011 OAC can therefore be defined as:

- Having an open and transparent methodology that can be easily reproduced by other researchers and be critically evaluated.
- Using open source programs, such as 'R', wherever possible, and publishing supporting documentation (including code) to allow for easier adaptation and reproduction.
- Soliciting the views of potential users to ascertain their priorities for a new classification.
- Constructing a classification that, where possible, reflects the general consensus of potential users and is therefore of the greatest use to the greatest number of people.

Adhering to these key concepts means the 2011 OAC methodology can be used as a template to create other open geodemographic classifications as required. If the 2011 OAC is deemed not to meet the needs of a particular application, then developers can utilise the same methodology, changing only the aspects relevant to their particular task, and create a bespoke classification. Although the 2001 OAC had an open methodology, it was more restrictive as it was produced in SPSS (a commercial statistical package). The advances in the number and quality of open source programs means that R can be used instead of SPSS and Quantum GIS instead of ESRI's ArcGIS.

The key concepts that drive the 2011 OAC can be viewed as a mixture of traditional geodemographic theory and outcomes of consultation. As discussed in Chapter 3, there are certain issues that are unique to the creation of the 2011 OAC, with the status of any future UK Census and the growth of alternative Open Data sources causing uncertainty in the development of the classification. The extent to which the 2011 OAC should embrace alternative data sources, or rely on Census data is not a choice that should be made in isolation. The decision on what data to use for the classifications forms perhaps the most important aspect of the 2011 OAC user engagement as the choice of data used impacts all aspects of the classification.

4.3. User Engagement on the 2011 OAC

The release of 2011 UK Census data provided the catalyst for the 2011 OAC to be developed. The Census has been the primary data source for geodemographic classifications for forty years, and the latest release is a chance to evaluate the continued use of Census data for this purpose in tandem with creating a new geodemographic classification. The choice of data sources may be the most important consideration, but there are many other factors that impact the usefulness and longevity of any classification. It was therefore important to design the 2011 OAC for general purpose applications across a wide target audience. To better understand user requirements the ONS and UCL conducted a user engagement aimed at individuals and companies with an interest in small area classifications, and the 2001 OAC in particular. The results of this are reported in Section 4.3.3.

4.3.1. Designing the User Engagement

Constructing the 2011 OAC user engagement required the selection of an appropriate method for extracting information from the desired audience. In the ONS User Engagement Strategy (ONS, 2010a) reference is made to focus groups, surveys and one to one meetings as methods of consulting with key stakeholders. Furthermore, there is extensive academic literature exploring the various approaches to qualitative data collection, with examples of user engagements employing a wide range of methods, including surveys, interviews, focus groups and participant observation. One example detailed in O'Brien and Toms (2008) utilises semi-structured interviews to "delve into the thoughts, behaviours, and feelings" (p. 941) of participants who were using online applications. Interviews and indeed focus groups provide obvious advantages as methods of user engagement, allowing exploration of the user's experiences by providing the opportunity to probe respondents to elaborate on their answers. However, these forms of data collection can be both costly and time consuming.

Examining the various methods available, it was decided that online self-completion questionnaires would be the most appropriate means of collecting the required information for the 2011 OAC user engagement. This involved the respondent answering a series of open and closed questions by themselves and then returning the completed questionnaire by e-mail.

The method was selected as it was cheap to administer and practical considering the respondents were geographically dispersed. It was also more useful for the respondents as they were able to complete the questionnaires at their own convenience. Questionnaires eliminate issues relating to an interviewer's presence, such as social desirability bias, by which the users may frame their responses in a way they consider to be more pleasing to the interviewer (Bryman, 2008). Finally, the efficiency of the method makes it an attractive option for data collection as numerous e-mails linking to the questionnaire could be sent out at the same time. Consequently a sufficient response rate meant a consensus could be drawn and therefore the information could be utilised in the design on the 2011 OAC.

The broad objective of understanding user requirements for the 2011 OAC, and specific details such as what data to use meant careful consideration of how the questionnaire was formed. The questions needed to be formulated in a way to reveal quantitative or qualitative information, while also adhering to a simple structure (Dixon and Leach, 1978). Although Parfitt (1997) stresses the content of a questionnaire should concentrate upon answering the formulated aims, no question should rely on recall of events too distant in the past. This meant refraining from asking questions about academic geodemographics prior to the 2001 OAC. It was felt respondents knowledge of historical geodemographic applications would be more limited, and of little relevance for the 2011 OAC.

Dixon and Leach (1978) offer other general advice in designing a questionnaire, such as using simple, short questions with terminology that is easy to understand. There should also be a combination of open and closed questions, with questions being unambiguous to ensure respondents are referring to the same thing. Dixon and Leach (1978) also explore the subject of how questions are perceived, with the attitudes of an individual being complex and unable to escape external influencing factors. As such, questions should actively encourage what could be termed an 'accurate' response, with one method of achieving this through the use of a Likert scale (Likert, 1932). This allows the extent to which a person agrees with a statement to be gauged on a scale, rather than using an opened ended text response.

The format chosen for the 2011 OAC user engagement questionnaire was a mixture of closed and open questions. This allowed for both quantitative and qualitative answers to

be given to better understand the thoughts, expectations and requirements of the new classification. Twenty questions were formulated which were then divided into five topic groups. These topic groups can be summarised as:

- Determining the respondent's understanding of the 2001 OAC and the extent to which they or their organisations use or used it. This could be expanded to determine how useful respondents find the 2001 OAC's structure, if they use any alternative geodemographic classifications and, if so, how they use them. This then helps to assess the priorities for a new small area classification (the 2011 OAC), and what its primary purpose should be.
- Determining what outputs users would find to be the most helpful for the 2011 OAC, and whether the classification should remain focused on delivering national coverage or be split into regions.
- Determining the format(s) that should be used to disseminate the 2011 OAC's outputs and how the classification should be presented.
- Determining at what spatial level the 2011 OAC should be constructed, and what data source(s) should be used. This then leads on to assessing how often the 2011 OAC should be updated (if this proves to be a possibility) in the future.
- Allowing for additional comments regarding the 2011 OAC from respondents to ensure pertinent information relating to the new classification can be recorded.

Pilot questionnaires are an important component of the design phase (Haring and Lounsbury, 1975; Dixon and Leach, 1978; Parfitt, 1997). They allow for different methods and wordings to be tested and to ensure the meaning of the questions are fully understood (Dixon and Leach, 1978). Ultimately the questions being asked will be meaningful if they have useful answers (Parsons and Knight, 2005). A pilot version of the 2011 OAC user engagement was distributed to the ONS and Keith Dugmore, Director of the Demographics User Group (DUG). A selection of DUG members were given the opportunity to assess the suitability of the questionnaire before it was made more widely available. The feedback received from this was favourable, with one example from Barclay's stating: "I've run through the draft and it seems very comprehensive, should work very well. I think the questions are meaningful and will give useful answers". The responses from the pilot study indicated that no changes needed to be made to the 2011 OAC user engagement. As such, it was made publically available for a six-week period from the 17th February 2012 to the 30th March 2012 (ONS, 2012j).

It was desirable to achieve as many responses as possible from a range of professional backgrounds. Haring and Lounsbury (1975) suggest coverage of the total research area is ideal; in this instance this would involve contacting every individual that has previously used the 2001 OAC. However, Parfitt (1997) has a more realistic view of surveys, suggesting they can be conducted with only a sample of respondents from the target population. As such, for the purpose of this study it was decided that a sample of the users of the 2001 OAC would provide a sufficient representation of the group as a whole, and while there is no general rule for devising this (Haring and Lounsbury, 1975) several factors such as the type and accuracy of data obtained had to be considered. Dixon and Leach (1978) observe that the larger the response the more meaningful the conclusions drawn from the data will be. Parfitt (1997) recognises this and admits with sampling there is a link between decreasing the sample size and increasing the sampling error; nonetheless it is still possible to draw meaningful conclusions from a small sample. In the case of the 2011 OAC user engagement the types of questions asked limited the total number of responses. If it had been solely a quantitative study then the aim would have been to achieve a large number of responses from randomly selected respondents. If it had been solely a qualitative study then the aim would have been to achieve a smaller number of in depth responses from a few selected respondents.

The 2011 OAC user engagement was promoted via specific websites and mailing lists such as the OAC User Group, Census Dissemination Unit, EDINA UKBORDERS and the retail industry Demographics User Group. This allowed for a larger number of individuals and companies with an interest in the 2001 OAC and geodemographics to respond, without limiting the user engagement to only a few preselected respondents. A sample response form can be found in Appendix A.

4.3.2. Findings from the User Engagement

The findings in this section are based on the summary of responses to the 2011 OAC user engagement published by the ONS in May 2012 (ONS, 2012k). The ONS and UCL received 38 responses from a mixture of local and central government, primary care trusts, other public sector organisations, consultancies, commercial organisations and academics. The findings of the 2011 OAC user engagement are summarised in this section. Copies of the fully tabulated results and some comments from respondents can be found in Appendix B. The respondent types have been classified into one of six groups as shown in Table 4.1.

Table 4.1: Responses by stakeholder group

Respondent Type	Responses	Percentage of total
Local Authorities (LA)	19	50
Central Government (CG)	2	5
Health (H)	3	8
Other Public Sector (PS)	3	8
Commercial Organisations & Individuals (CO)	7	18
Academia (A)	4	11
Total (All)	**38**	**100**

From the 38 responses half came from local authorities and 12 asked for their comments to remain confidential. The responses from various stakeholder groups reflect the continued interest in the 2001 OAC and how important it is for the 2011 OAC to cater to number of different disciplines. The results and a brief interpretation for each question are detailed below. Note that figures are presented as percentages that may not sum exactly to unity due to rounding. Counts are included in the brackets.

4.3.2.1. The current 2001 Area Classification for Output Areas

Question 1: Do you know what the current 2001 Area Classification for Output Areas (2001 OAC) is?

Table 4.3: Responses to Question 1 of the 2011 OAC user engagement

	All	LA	CG	H	PS	CO	A
Yes	97 (37)	100 (19)	50 (1)	100 (3)	100 (3)	100 (7)	100 (4)
No	3 (1)	0 (0)	50 (1)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

Of the 38 responses received only 1 respondent did not know what the 2001 OAC was. The 2001 OAC can therefore be considered to have had wide penetration across multiple stakeholder groups that have an interest in geodemographic classifications. It is unsurprising that the majority of respondents were familiar with the classification as the 2011 OAC user engagement was primarily aimed at 2001 OAC users.

Question 2: Do you (or your organisation) currently use the 2001 OAC?

Table 4.4: Responses to Question 2 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
Yes	61 (23)	47 (19)	0 (1)	100 (3)	67 (3)	71 (7)	100 (4)
No	39 (15)	53 (0)	100 (1)	0 (0)	33 (0)	29 (0)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

Twenty-three of the respondents indicated they, or the organisation they work for, currently use the 2001 OAC. The proportion of organisations using the 2001 OAC varied depending on stakeholder group; over half of the local authorities that responded do not currently use the 2001 OAC, while over half of commercial organisations and individuals do use it. The 23 respondents who indicated that they or their organisation currently use the 2001 OAC were also asked an additional question:

2a. If you answered "Yes" how long have you (or your organisation) been using the 2001 OAC for?

The majority (87%) of those currently using the 2001 OAC have been doing so for 2 years or more. This is a pattern seen in all stakeholder groups that currently use the 2001 OAC. These historical users of the 2001 OAC can be considered to have a good understanding of what their requirements for the 2011 OAC are.

Question 3: If you answered "No" to Question 2 have you (or your organisation) previously used the 2001 OAC?

Only the 15 respondents who indicated they, or their organisations, do not currently use the 2001 OAC were eligible to answer this question. 67% of users did however indicate they had used the 2001 OAC previously. This question was then subdivided to elicit further information:

3a. If you answered "No" why have you never used the 2001 OAC?

3b. If you answered "Yes" how long ago did you (or your organisation) stop using the 2001 OAC?

3c. If you answered "Yes" why did you stop using the 2001 OAC?

These questions sought to try and find why 15 of the respondents were not currently using the 2001 OAC. Respondent 20 from central government who had never used the 2001 OAC said: "there was a lack of general awareness that such a free product produced centrally existed". This raises the issue of publicising the existence of free open-source geodemographic classifications.

Ten respondents who had previously used the 2001 OAC indicated that the main reason they had ceased to use the classification was due to a shift towards using only commercially available systems. Respondent 5 commented: "we currently hold a Mosaic licence that offers data at a more local level", while respondent 7 indicated that the 2001 OAC "groups and descriptions do not reflect reality". The outcome is the respondents who use commercial systems do so, not only because they offer services that the 2001 OAC does not, but because the 2001 OAC is believed to be flawed, or even wrong, in some areas of the UK.

Question 4: What alternative commercial geodemographic classifications do you (or your organisation) use?

The responses from all stakeholders indicated that commercial geodemographic classifications are heavily used; the dominant classifications being ACORN by CACI and Mosaic by Experian (both used by 31% of respondents). Several respondents gave reasons for this: "the commercial product gives better discrimination for smaller geographic areas and has proved to be more accurate than [the 2001] OAC, particularly in rural areas" [respondent 8], "Mosaic is user friendly" [respondent 22] and "Mosaic provides a household and postcode-level granularity and reflects the diversity of the area far better than [the 2001] OAC" [respondent 15]. Another comment, that commercial products "provide better discrimination than [the 2001] OAC and have less issues for London" did also note: "however, they are closed source" [respondent 37], acknowledging one potential benefit of the 2001 OAC.

There were 10 stakeholder respondents who stated they did not use commercially available classifications, which indicates their continued use of the 2001 OAC. Respondent 2 justified this on the grounds that "[the 2001] OAC offers much better value, basis & user engagement," while respondent 9 stated that they "have no budget for commercial products".

Question 5: Please indicate the geographical coverage(s) you favour when using a geodemographic classification?

This question attempted to understand how different users use geodemographic classifications, and if they require a countrywide product or one that focuses on a smaller geographic region. The variation seen in the responses from the different stakeholder groups provides an indication of the different ways geodemographic classifications can be used. For local authorities the preferred geographical coverage is at a more local level, while central government requires a more national view. There was no consensus from the respondents, with each geographical coverage option gaining multiple responses. This in part reflects the personal needs and requirements of each respondent, but this variation in responses indicates a demand for a more flexible approach when constructing the 2011 OAC to cater for a larger percentage of user needs. Some users found the UK-wide coverage of the 2001 OAC to be a disadvantage as it meant they obtained less detailed information relating to their specific local area. On the other hand, some users saw this as an advantage, appreciating the opportunity to compare their locality to others across the country.

Question 6: Would you welcome a new version of the 2001 OAC?

Table 4.5: Responses to Question 6 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
Yes	89 (34)	95 (18)	100 (2)	67 (2)	100 (3)	86 (6)	75 (3)
No	8 (3)	0 (0)	0 (0)	33 (1)	0 (0)	14 (1)	25 (1)
No Answer	3 (1)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

There was a general consensus across all stakeholder groups and a clear majority of respondents (89%) that a 2011 Area Classification for Output Areas (the 2011 OAC) would be welcomed. The number of respondents who would welcome the 2011 OAC is greater than those who currently use the 2001 OAC. Caution should however be attached to this figure as welcoming the creation of the 2011 OAC does not equate to the same number of people or organisations using it. It does however indicate that the majority of respondents do see creating the 2011 OAC a worthwhile endeavour.

Question 7: Should a new 2011 Area Classification for Output Areas (2011 OAC) be a general purpose classification (like the current 2001 OAC), or should it focus on producing specialised variants (such as health, education or crime)?

Table 4.6: Responses to Question 7 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
General purpose	55 (21)	53 (10)	100 (2)	33 (1)	33 (1)	57 (4)	75 (3)
Specialised variants	45 (17)	47 (9)	0 (0)	67 (2)	67 (2)	43 (3)	25 (1)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

The divide in opinion shown in the responses to this question reflects the split between those seeking a general purpose classification and those wanting a niche solution. Respondent 34 commented that:

> *"The idea of differing versions of the [2011] OAC is an interesting one. I think a key requirement of the main [2011] OAC and any alternative version is clarity about what is being measured. There is a tendency for general classifications to try and cover a wide range of variables, meaning that the final classification tries to cover too many bases and loses clarity. Thus a series of specific focussed sub classifications might be a valuable approach."*

Like the responses for geographic coverage (question 5), there does seem to be a demand for a more flexible approach when constructing the 2011 OAC in order to cater to different needs.

Question 8: Flexibility in specifying the variables that are to make up the 2011 OAC would open up a range of options for area classification using Open Government Data. Is it important to you that the 2011 OAC be directly comparable – in terms of similar Census data being used to construct it - with the 2001 OAC?

Table 4.7: Responses to Question 8 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
Yes	13 (5)	11 (2)	0 (0)	67 (2)	0 (0)	0 (0)	25 (1)
No	87 (33)	89 (17)	100 (2)	33 (1)	100 (3)	100 (7)	75 (3)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

The majority of respondents indicated that they do not believe the 2011 OAC needs to be directly comparable with the 2001 OAC. Two of the three respondents from the Health stakeholder group felt the two OAC's should be comparable. Respondent 26 did comment: "although answering 'No', for continuity and 'what has changed?' purposes direct comparability to 2001 would help". Another respondent stated: "a stable classification would be more useful. A classification stable 1991-2011 ideally" [respondent 35]. The creation of a classification that spans numerous Censuses would no doubt be a useful product; however, this would go beyond the specific aims of this project. Although, an open and transparent approach in the construction of the 2011 OAC does allow users the possibility of modifying the methodology for bespoke purposes, including adding a temporal component to a classification. The 33 respondents who said 'No' were then asked a follow up question to ascertain what alternative priorities they had:

8a. If you answered 'No' then what are the other priorities that are important to you in the construction of the 2011 OAC?

The dominate answers given to this question were 'updateable' (36% of responses) and 'better variables' (46% of responses). These responses suggest potential users of the 2011 OAC value the importance of how a classification is composed. Respondent 8 commented: "I would like to see a wide range of variables being used from different sources (Open Government Data)". Using such variables that could also be updated would seem to adhere to what a large number of respondents requested. Only 2 respondents indicated they would consider the use of Open Data to be a priority for use with the 2011 OAC. This would seemingly contradict the desire to create an updateable classification, although it can be thought as the respondents suggesting they would value an updateable classification and they do not mind where the data come from or what label this data are given. There were other suggestions from respondents: "better distinctions between the groups and names" [respondent 10] and "more documentation

that is accessible rather than technical on each cluster" [respondent 31] suggest the role in assigning group names and providing additional documentation is also considered important.

Question 9: The 2001 OAC divides the population of the UK into 7 Supergroups, 21 Groups and 52 Subgroups. How would you describe this framework when using the 2001 OAC for your particular purposes?

The responses to this question indicate overall the respondents felt the way the 2001 OAC divided the UK was either 'satisfactory' or 'good', with 75% of the responses indicating this. Only a single respondent felt the structure was excellent and a further 7 found it either limited or extremely limited. This links with the previous questions that indicates respondents have different priorities from a geodemographic classification. One respondent commented that they only used the 7 Supergroups and 21 Groups as they find the 52 Subgroups too detailed for their particular purposes.

4.3.2.2. New for the 2011 Area Classification for Output Areas

Question 10: Thinking about how you use and interpret the 2001 OAC, how useful do you think each to the following options would be to you for the 2011 OAC? (1 = Not at all useful to 5 = Extremely useful)

The options given were:

10a. Maps in PDF (or similar) format that are not interactive

10b. Online interactive maps with clickable details

10c. Mapping against different backdrops (such as Google Maps or OpenStreetMap)

10d. Correlation tables (showing to what extent the variables within the classification correlate with each other)

10e. Bar graphs of the group's attributes

10f. Radial plots of the group's attributes

Full tabular results to this question are shown in Section B.1.1.2 in Appendix B. The responses from these questions indicate that across all stakeholder groups there is a need to understand what the classification means, and not just show what the classification is. For example, 74% of respondents rated the inclusion of bar graphs of a

clusters attribute as 'useful' or 'extremely useful', and 56% said the same for radial plots. These outputs allow for a better understanding of the composition of the groups produced, rather than just looking at the visual distribution that a PDF map offers (something 21% of respondents rated as 'not at all useful'). These type of responses differ from question 8 where they indicated a greater interest in how the classification is composed rather than how it is presented. The responses to question 8 do not devalue the importance of understanding the 2011 OAC, and there was a particular interest in being able to view tables of how the variables used in the classification correlate with each other. Outputs such as these and bar graphs were not made available for the 2001 OAC and neither were any online interactive maps provided initially. The majority of respondents (76%) suggested they would find such a facility for the 2011 OAC to be either 'useful' or 'extremely useful' and would value the ability to overlay layers such as Google Maps or OpenStreetMap (OSM).

Question 11: Thinking about your own understanding of the existing 2001 OAC, how useful do you think each of the following options would be to you for the 2011 OAC? (1 = Not at all useful to 5 = Extremely useful)
The options given were:

11a. Group Name
11b. Graphical Representation (radial plots and bar graphs)
11c. Group definitions (a written summary of the key characteristics of each group)
11d. Key points of characteristics you would expect to find in each group
11e. Written 'pen portraits' of typical households found within each group
11f. Written 'pen portraits' of typical housing and built environments found in each group

Full tabular results to this question are shown in Section B.1.1.2 in Appendix B. Following on the theme of question 10 of understanding the classification, all the options given in question 11 on the different methods that could be used to enhance understanding of the 2011 OAC were popular across the stakeholder groups. The majority of respondents (87%) identified each cluster having a name to be either 'useful' or 'extremely useful'. Written descriptions were also considered particularly favourable, with 87% of respondents stating pen portraits would be 'useful' or 'extremely useful' for aiding understanding household characteristics. In addition to this, 79% of respondents

indicated that pen portraits of physical environment would be 'useful' or 'extremely useful', and 97% said the same about knowing what the key characteristics of each group were. These responses suggest a need for more detailed understanding of the classification to compliment the quicker interpretation offered by a cluster name, bar graph or radial plot.

Question 12: Do you agree with the view that it would be helpful to adjust the composition of each group for different parts of the UK (so, for example, there might be separate classifications made for London, or Scotland)?

There was no consensus from the respondents regarding if separate classifications should be part of the 2011 OAC. Across the stakeholder group there were slight variations; 42% of local authorities stated they 'agreed' with the concept compared to a 75% of academics who either 'disagreed' or 'strongly disagreed' with it. Only a limited number of respondents explained their decision, but one comment: "the downside is losing comparability between areas" [respondent 3] demonstrated an understanding of the problem of having multiple separate classifications. Other comments from respondents indicated that a national 2011 OAC along with separate classifications should be constructed: "I would welcome local variants, but these should supplement rather than replace the UK-wide OAC" [respondent 29] and another suggestion: "we would prefer to have consistent classifications across GB/UK and finer region specific subcategories, which could be aggregated up to the consistent GB/UK classifications, may be useful" [respondent 20]. This again suggests the need for the construction of the 2011 OAC to allow for potential variants created by/for potential users.

Question 13: Please identify what, if any, extra features would you like the 2011 OAC to have when compared with the 2001 OAC

A large number of respondents did not answer this question. Of the 20 that did, their responses indicate, like questions 10 and 11, a greater understanding of the groups would be welcome with the 2011 OAC. The comments some respondents made are a clear indication of this: "I like the idea of Pen Portraits as this enables people to understand the groups better" [respondent 7], "Group and Type names in particular from the outset, also pen portraits and an interactive multimedia guide with visualisations of data variables" [respondent 4], "it would be good to have names for the 52 sub-groups… a name is easier to explain than a number" [respondent 18] and "the difficulty with the 2001 OAC was that the characteristics even of the Supergroups weren't that clear, and were worse with the smaller groups. I think the Subgroups are

possibly superfluous. Clear names and descriptions are needed for every categorisation" [respondent 34]. Pen Portraits, detailed descriptions of each group's characteristics, in particular appear to be desirable to local authorities, central government and commercial organisations. There were other extra features identified but these varied more by the respondent's particular needs. An example of this is respondents wished to have features that are currently only available in commercial systems, such as a postcode level classification. There was also the generic comment made by multiple respondents that the 2011 OAC should aim "to be more like Mosaic".

4.3.2.3. Dissemination of the 2011 Area Classification for Output Areas

Question 14: Which methods of dissemination for the 2011 OAC would you be most likely to use?

Respondents were allowed to select any combination of the following five options:

- ***Online interactive mapping***
- ***Enhanced online interactive mapping***
- ***Microsoft Excel/CSV file(s)***
- ***Software to append the 2011 OAC codes to postcodes***
- ***Digital Boundary Data***

Full tabular results to this question are shown in Section B.1.1.3 in Appendix B. There was limited variation in the responses, suggesting while stakeholders may use geodemographic classifications for different purposes, they still rely on similar outputs. The majority of the 19 local authorities who responded said they would most likely use CSV files or digital boundary data. Other stakeholders provided similar responses making CSV files and digital boundary data the most popular method of dissemination. Other methods, such as online and enhanced online interactive mapping generated only limited interest when compared to the other dissemination options. Making multiple outputs available would seem the most appropriate way to satisfy these identified needs.

Question 15: Other data sources could be used to give greater context to the 2011 OAC. Rather than contributing to the classification itself, these could be used to help visualise the 2011 OAC in different ways. What (if any) data sources would you like to be able to use alongside the final 2011 OAC output?

Respondents were allowed to select any combination of the following seven options:

- ***Index of Multiple Deprivation (IMD)***
- ***Temporal data***
- ***Health related data***
- ***Land use data***
- ***Weather history data***
- ***Travel to Work areas***
- ***Other survey data***

Full tabular results to this question are shown in Section B.1.1.3 in Appendix B. The IMD was the most desired additional data source to aid the visualising of the 2011 OAC, with 33 respondents indicating this. The use of the IMD would however not be consistent across the UK due to each country producing their own variant. Other data sources, such as those related to health, had less support. The results varied depending on the requirements of each stakeholder group, with must options being selected by a limited number of respondents. Aside from the IMD, there appears to be limited desire to use additional data sources in order to help visualise and add greater context to the 2011 OAC.

4.3.2.4. Construction of the 2011 Area Classification for Output Areas

Question 16: There are multiple levels of spatial resolution that data can be produced. In addition to Output Areas are there any other spatial resolutions you believe would benefit from having their own classification?

Full tabular results to this question are shown in Section B.1.1.4 in Appendix B. Lower Layer Super Output Areas (LSOAs), Data Zones (DZs) and Super Output Areas (SOAs) had the most responses to this question, with 32 of the respondents indicating these spatial levels would benefit from having their own classification. The concept of a classification at local authority level was also popular with 18 respondents indicating they would find such a product beneficial. Three respondents, all from commercial organisations suggested creating a classification as postcode level, to emulate those offered by companies such as Experian and CACI. This is an understandable, but difficult, request to fulfil when the 2011 OAC, like its predecessor, focusses on data at the OA level. OAs are the smallest level at which Census data are produced and typically are made up of

multiple postcodes. It is important that the 2011 OAC does not lose focus and as such "OAs remain top priority - the others are nice to have" [respondent 33]. This question does re-affirm earlier conclusions that the respondents all use geodemographic classifications differently and therefore have different requirements for what would be a 'perfect' 2011 OAC for them.

Question 17: The 2001 OAC uses only 2001 Census data in its construction. It has been suggested that, in addition to using 2011 Census data, it might be possible for the 2011 OAC to be enhanced with supplementary non-Census Open Data sources, and updated periodically over time. Would you find this beneficial?

Table 4.8: Responses to Question 17 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
Yes	87 (33)	95 (18)	0 (0)	100 (3)	100 (3)	100 (7)	50 (2)
No	5 (2)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	25 (1)
Do not know	8 (3)	0 (0)	100 (2)	0 (0)	0 (0)	0 (0)	25 (1)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

The majority of respondents (87%) indicated they would find it beneficial for the 2011 OAC to be updated using non-Census Open Data sources. There were however varying degrees of enthusiasm in comments left by the respondents. Some were positive about updating the 2011 OAC:

- "This brings the Census-based classification to life!" [respondent 10]
- "Would enable the classification to more closely reflect current conditions by picking up changes such as new housing developments." [respondent 1]
- "To address the risk that [the 2011] OAC is perceived as irrelevant or out of date - periodic updates bring [the 2011] OAC in line with more commercial products and demonstrate on-going relevance (as well as show change over time)" [respondent 2]
- "The addition of more data will enhance the value of the [2011] OAC and using Open data will promote the aims of the open initiative further." [respondent 28]
- "I think constantly adding new analysis to the [2011] OAC categories keeps them fresh in peoples' minds and constantly adds new understanding" [respondent 7]

- “Enriches and updates the 2011 Census results - as long as no cost involved” [respondent 9]
- “This would be useful where there are gaps in the Census questions coverage.” [respondent 6]

There were however some respondents who took a more cautious line, while some disagreed with any sort of updating:

- “Updating may give [the 2011] OAC more credibility with people liable to dismiss Census data as too out-of-date to be relevant. However, when using the 2001 OAC with groups more than 10 years after the Census, I've found that most recognise the designation of their neighbourhoods - the purpose of updates should therefore be more than "enhancement.” [respondent 29]
- “Extra data yes; updating no (because Census not updatable in the same way)” [respondent 36]
- “The addition of non-Census data will impact on reproducibility - unless the sources are entirely Open Data; which at OA level is unlikely” [respondent 37]
- “I think there is a role for a Census only classification” [respondent 19]
- “Depends on the types of dataset used and how they could enhance the classification.” [respondent 21]

The range of different opinions the respondents had on this issue is perhaps a reflection of the varying knowledge the respondents have about geodemographic classifications. Some respondents indicated what they would like without knowing what is technically possible, while other respondents have a better understanding of the limitations of geodemographics and data and had different opinions as a result. The conflicting requirements that arise from the answers to this question cannot be easily resolved, and the final choice of data source(s) ultimately must provide the greatest benefit to the greatest number of people and/or organisations.

Question 18: It is unlikely that many Open Data sources will offer UK wide coverage. What extent of coverage do you believe is a minimum requirement for an acceptable general purpose classification?

Table 4.9: Responses to Question 18 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
UK only	10 (4)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	75 (3)
Countrywide	43 (18)	43 (9)	50 (1)	0 (0)	40 (2)	71 (5)	25 (1)
Regional	17 (7)	19 (4)	0 (0)	0 (0)	20 (1)	29 (2)	0 (0)
Local Authority	26 (11)	33 (7)	0 (0)	100 (3)	20 (1)	0 (0)	0 (0)
Ward	2 (1)	0 (0)	0 (0)	0 (0)	20 (1)	0 (0)	0 (0)
No Answer	2 (1)	0 (0)	50 (1)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (42)**	**100 (21)**	**100 (2)**	**100 (3)**	**100 (5)**	**100 (7)**	**100 (4)**

The responses again reflect the different needs of respondents. Depending on what the respondent, or their organisation, uses geodemographic classifications for will influence the coverage they believe is a minimum requirement for a general purpose classification. The two most common responses, countrywide and local authority, would suggest that a classification created using Open Data would still need to allow comparison across a country and provide detailed characteristics at the local authority level. At present such a scenario would be challenging to create, due to the non-availability of appropriate data sources. Even it if were feasible it would still require national level classifications supplemented by separate ones for every local authority. Any benefits achieved from utilising Open Data sources would outweighed by the negatives such a scenario would create.

Question 19: If the 2011 OAC could be updated with new data, how frequently should this be done?

Respondents were allowed to select one of the following three options:

- ***Once a year***
- ***Every two years***
- ***Every three years or longer***

Full tabular results to this question are shown in Section B.1.1.4 in Appendix B. The responses to this question suggest that stakeholders would welcome updating of the 2011 OAC, if possible, on a regular basis. The preferred choice was once a year, with 42% of the responses with every two years having 26% and every three years or longer having 29%. Within this there could be an element of desire on the part of the

respondents, rather than any particular need they may have identified. The actual need for a classification to be updated regularly is debateable (as discussed in Section 2.8.3), but it does not help encourage continued use if it is perceived to be out of date. Regular updating of the 2011 OAC in some form would be both beneficial to the how the classification is perceived and welcomed by the majority of respondents.

Question 20: Change in the social, economic and demographic structure of areas in the UK occurs at different rates. Instead of updating the 2011 OAC it might be possible to use non-Census sources to flag areas where population changes have occurred, enabling the user to recognise parts of the UK where the classification had probably become unreliable. Would you find this helpful?

Table 4.10: Responses to Question 20 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
Yes	97 (37)	95 (18)	100 (2)	100 (3)	100 (3)	100 (7)	100 (4)
No	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Do not know	3 (1)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

With the exception of 1 respondent who did not answer, the 37 other respondents all indicated they would welcome some form of uncertainty measure to be included as part of the 2011 OAC. Respondent 34 commented: "I think that a system that flags areas that have changed would be a useful product in its own right". There were however some more cautious comments made by another respondent: "if this proposal were undertaken to flag where the segmentation has become unreliable the commercial segmentation vendors would exploit it" [respondent 30]. They did however offer an alternative suggestion: "social media use by [the 2011] OAC could be an interesting avenue of exploration given the increasing volume of freely available geocoded social media" [respondent 30]. The responses, along with the comments made, suggest that such a measure is desired by the respondents, but should be approached with caution.

4.3.2.5. Other Comments

The respondents were also given an opportunity to leave any additional comments relating the 2011 OAC. A selection of these comments have been included in Appendix B. These specific comments were used in tandem with the tabulated findings of the 2011 OAC user engagement to help define what the user requirements for the classification were. These are reported in Section 4.4.

4.4. 2011 OAC Findings and User Requirements

The responses to the 2011 OAC user engagement provided detailed insight into what potential users of the new classification considered important issues. Although there was a reasonable response rate to the user engagement, the variation between the respondent's backgrounds made it difficult to create a clear set of user requirements. Questions where general agreement occurred tended to be statements of intention, such as 89% of respondents indicating they would welcome the 2011 OAC. On more specific issues, such as which data source(s) to use, there was less agreement. Instances such as these were a reflection that respondents utilise geodemographic classifications in different ways. As a result, user requirements and expectations appeared limited to this outlook. Although there was not total agreement on many issues, there were still some general themes that could be found in the responses. These themes can be broken into the following six points:

- Using the best possible data source(s) for the 2011 OAC
- Open Data to have a role with the 2011 OAC
- The need to evaluate the effectiveness of the 2011 OAC
- The 2011 OAC to be a general purpose geodemographic classification
- Provide additional information about the outputs of the 2011 OAC
- The need to publicise the 2011 OAC

These points form the basis of the user requirements of the 2011 OAC and are explored in more detail below. Each point was carefully considered in order to select the best options for the greatest number of respondents.

4.4.1. Using the best possible data source(s) for the 2011 OAC

The respondents did not agree on what data should be used to create the 2011 OAC. With no general consensus formed it was decided that the 2011 OAC would focus on using only Census data. The rationale behind this decision was based on the desire for the 2011 OAC to be created at the smallest spatial level, and at present the Census remains the only UK-wide data source that provides data at a small enough granularity to accomplish this. As such the 2011 UK Census can be considered the best possible data source for the 2011 OAC and supplementing it with additional Open Data sources introduces too many uncertainties with little benefit.

4.4.2. Open Data to have a role with the 2011 OAC

The majority of respondents who expressed a desire to use Open Data sources did so because they felt the 2011 OAC should be periodically updated. This was considered one of the priorities for the new classification (along with selecting better variables). The use of only Census data to construct the 2011 OAC means it will not be possible to replicate the methodologies used by commercial operators to update their classifications as they incorporate a range of data sources. The alternative proposal, of including an uncertainty measure with the 2011 OAC, to identify OAs and SAs as having potentially undergone some change in either their social or built up environment, thereby suggesting their geodemographic assignment may be uncertain, was welcomed by all but one respondent. Although this is not a direct replacement for continual updates of the classification, it would allow the continued relevance of the 2011 OAC to be assessed.

4.4.3. The 2011 OAC as a general purpose geodemographic classification

The majority of respondents wished for the 2011 OAC to be a general purpose classification, but not directly comparable with the 2001 OAC. A provision to allow for specialised variants was stated as a desirable feature, both to allow for regional classifications and bespoke classifications (e.g. health). This was indicative of a desire expressed by respondents to have more control on adapting the classification to their own purposes. This relates back to a key concept of the 2011 OAC to have an open and transparent methodology. Making resources such as code available so users can modify the classification for their own purposes means this will be a possibility.

4.4.4. The need to evaluate the effectiveness of the 2011 OAC

Some former users of the 2001 OAC felt that the classification groupings and descriptions were flawed and not a reflection of reality. No geodemographic classification will ever be perfect, but the advantage of the 2011 OAC being constructed within an academic framework means that it can be critiqued in an open and transparent way with all ground-truthing exercises documented. Any flaws that are found can be highlighted to the wider user base and fixed. The issues the 2001 OAC had with London were highlighted by some respondents, noting that closed commercial systems offered better discrimination. The creation of the 2011 OAC is not a guarantee that such issues will not be repeated, but the key concept of creating an open and transparent methodology means that the classification can be re-engineered by users if desired.

4.4.5. Provide additional information about the outputs of the 2011 OAC

The respondents generally agreed that the 2001 OAC divided the UK in either a satisfactory or good manner with 7 Supergroups, 21 Groups and 52 Subgroups. Some respondents indicated they would welcome a greater understanding of these groups and in particular naming the Subgroups and descriptions for all of the groups. Although these were the most favoured features of the 2011 OAC, they was a general desire to make multiple outputs available that explained the composition of the classification.

4.4.6. The need to publicise the 2011 OAC

The awareness of the existence of a free open geodemographic classification was high amongst respondents, with only one respondent unaware of the 2001 OAC. Although this lack of awareness was in the minority, the targeting of the 2011 OAC user engagement at past and present 2001 OAC users meant that these responses were unsurprising. The respondent who was not aware of the 2001 OAC did not cite any specific reason for their general lack of awareness, but this does suggest a need to publicise the existence of the 2011 OAC beyond current and past 2001 OAC users. The 2011 OAC should not be limited to being only for these users, and the classifications wider adaption will be important to its continued use in the future.

4.5. Conclusions

The formation of the 2011 OAC shares many similarities with previous geodemographic classifications, yet can be considered a unique product. This uniqueness is derived from a combination of key concepts and the user requirements of the classification, which, acting together provided a blueprint on the processes involved in constructing the 2011 OAC. The key concepts are both a reflection of the decisions made internally and of external factors, such as the Open Data agenda, and are a natural by-product of these different demands. Perhaps the most important key concept was the desire to collect the views of potential 2011 OAC users to ascertain their requirements for the classification. There were many different methods available that could have been used to facilitate the collecting of these views, but an online questionnaire was selected as the most suitable method. The 2011 OAC user engagement ran for 6 weeks and received 38 completed replies, providing valuable insight into respondent's expectations of the new classification.

The responses to the 2011 OAC user engagement lacked any form of general consensus on multiple questions. Patterns emerged of respondents from different backgrounds having different priorities for the 2011 OAC, for example, those from the health stakeholder group indicated they wanted a more health-focused classification. Although variations existed in some of the specific expectations for the 2011 OAC, broader themes did emerge. The use of Open Data had conflicting responses, although these gave the impression that using the best data source(s) for the 2011 OAC should be a priority. As such, this formed one of the six user requirements, rather than specifically stating which data source(s) to use. The other five user requirements were selected as they either reflected the views of a number of respondents, or because they enhance the 2011 OAC and therefore appeal to a wider range of potential users.

The combination of the key concepts and user requirements provided a clear set of guidelines on the creation of the 2011 OAC. The semi-consultation led nature of this meant all steps required to create the 2011 OAC had some form of input from potential users. Looking at how this approach fits within the wider geodemographics field, the key concepts devised can be considered applicable to the creation of any open geodemographic classification. They set out clear principles on how a geodemographic classification can be created, with the consultation element and subsequent user requirements an extension of this. The user requirements for the 2011 OAC, while important to the creation of this particular classification, have to be considered unique.

They are a reflection of a certain set of expectations, based on the needs of individuals and companies, availability of resources and by what is technically achievable. Despite their transient nature, adhering to these expectations is what gives a classification on-going relevance and thereby increasing its appeal to a wider user base.

Chapter 5

Temporal and Spatial Stability of Small Area Classifications

5.1. Introduction

This chapter is based on Gale and Longley (2013) and examines geodemographic classifications, their inherent uncertainties and methods for coping with the problems that these uncertainties present. A large body of academic literature analyses the fuzzy nature of geodemographics (in other words, how accurate the assignment of a group is to a particular area), such as Fisher and Tate (In Press), but there is very little focus on the spatial and temporal uncertainty of classifications.

Commercial classifications use ancillary data sources to regularly update their products, which is a feature that is popular with users. This is not possible with academic geodemographic classifications built with Census data due to the unavailability of regularly updated data sources at fine levels of granularity. The 2001 OAC was released over five years after the 2001 Census Day (29th April 2001), and by the time the 2011 OAC is released, the representations it provides will be based on data over 13 years old. This may lead to the perception that it is out of date and inaccurate, which may lead many users to choose to abandon it (see Section 4.3.2).

Extending the longevity of current and future classifications, like the 2011 OAC, can be considered an important issue to users. The alternative to the commercial sector methods examined in this chapter is to create spatio-temporal uncertainty indicators from the limited amount of appropriate Open Data available. These indicators allow areas of significant change to be highlighted and question the validity of the original geodemographic assignment, giving users more flexibility in the utilisation of a classification.

Uncertainty indicators are based on the premise that change in the UK varies both geographically and over time. Examining the variances of change across the UK gives a general indication of how different areas exhibit change characteristics, both spatially and temporally. As only a limited number of data sources are currently available at the finest spatial level, temporal uncertainty indicators were solely created using mid-year population estimates (MYEs) and dwelling stock counts. The geodemographic coverage of these indicators varied; MYEs provided UK-wide coverage while dwelling stock counts could only be used for England and Wales.

A third indicator was created by combining the MYEs and dwelling stock counts; again this was only possible for England and Wales. These indicators were applied to the 2001 OAC between 2002 and 2010 to test their usefulness in highlighting the emergence and propagation of uncertainty during this period. The temporal variance in how uncertainty developed was examined by both the spatial variance within England and the other UK countries and by the 2001 OAC Supergroups.

The chapter concludes with looking at the inherent uncertainties that exist with the temporal uncertainty indicators themselves. Additionally, the continued development of small area data source availability is examined, and how this may lead to the creation of more comprehensive temporal uncertainty indicators in the future.

5.2. Uncertainty in Geodemographic Classifications

The provision of incremental updates is marketed as a key advantage to commercial geodemographic classifications. Ancillary sources are used to enrich and update the classifications, with users equating 'frequently updated' with being more accurate and the assumption that these options are the most useful; despite the origins of some of the ancillary sources remaining unknown (Experian, 2010; CACI, 2013b). Users can therefore view the lack of any inter-censual updates to classifications like the 2001 OAC as a negative, due to its sole reliance on data from the Census. Indeed, the lack of any updates during the lifetime of the 2001 OAC was a concern for some users of the classification, as shown by the responses to the 2011 OAC user engagement discussed in Chapter 4.

The enduring relevance of academic and commercial geodemographic products, at least from a user perspective, differs greatly. The 2001 OAC is now perceived to be of limited

use in areas of the country that have changed rapidly over the last decade; while commercial products which have been 'freshened up' using a range of sources that are of unknown provenance, continue to be popular. The clear difference in approaches between academic and commercial geodemographic classifications would appear to have had a detrimental impact on the long-term use of the 2001 OAC and future use of the 2011 OAC. Anecdotal evidence from past users suggests that they stopped using the 2001 OAC after a few years because they felt it was no longer accurate. Any attempt to update a classification like the 2001 OAC, or the 2011 OAC in the future is a complex task. Any geodemographic classification that uses Census data will encounter the problem of data latency. It is not possible to reproduce every aspect of a Census dataset without undertaking another Census, and towards the end of the decennial cycle the reliability of the data becomes increasingly uncertain. The infrequent nature of a national Census means that the traditional methods for temporal updates employed by the commercial sector cannot be applied to classifications like the 2001 OAC and the 2011 OAC, and it is necessary to consider alternative methods.

The increasing proliferation of government Open Data sources appears to offer some solutions (see Chapter 3). Yet despite improvements in the availability and dissemination of Open Data, the lack of datasets currently available at OA level is problematic. As classifications like the 2001 OAC and 2011 OAC are built at the OA and SA level, alternatives to traditional updating methods need to reflect this; recourse to coarser grained Open Data is therefore not ideal as local detail is lost. Additional complications arise out of the different data dissemination conventions in England and Wales, in Scotland and in Northern Ireland, which has implications on UK-wide classifications.

The only viable alternative at present to updating public sector data used in a classification's construction is to use the small number of measures obtainable at the OA level to construct temporal uncertainty indicators to highlight the stability of geodemographic assignments. These can subsequently be compared at national and regional scales, identifying areas in which significant changes in demographic compositions have occurred at the small area level. These are not designed to update the classification, but rather provide an indication of where updates are likely to be necessary.

Any area deemed to have experienced significant change over time is flagged as being potentially uncertain, and therefore a less reliable assignment. The 2001 OAC was the ideal classification to test this method. Changes over the past decade are identifiable, and the distinctive nature of how the temporal and spatial small area change across the UK impacted the applicability of the classification over a longer time period could be evaluated. The benefit of implementing this type of method is that it could be applied to any geodemographic classification.

During the 2011 OAC user engagement, participants were asked about their desire for this type of measure to be constructed (see Section 4.3.2.4); the majority of respondents indicated they would welcome such a facility. This aspiration, coupled with respondents indicating that they gravitated towards commercial geodemographic classification products because of the perception that classifications like the 2001 OAC are unreliable due to the absence of any temporal updating, suggests that the creation of a new classification like the 2011 OAC alone will be insufficient in altering this perception. In addition, the 2011 OAC by itself does not address the inherent issue of the spatially variable degradation in the reliability of certain geodemographic assignments over time. The creation of temporal uncertainty indicator(s) that could work alongside the 2011 OAC can therefore be seen as a positive development, which should work to reduce the perception that the classification will become irrelevant a few years after its release.

To date, this notion of temporal uncertainty has not been explored in any detail within the wider discourse of geodemographics: rather, the main focus of uncertainty within geodemographics has been at the initial stage of creation of a classification, specifically with respect to the cluster assignment procedure. For example, there are uncertainties inherent in the assignment of any area to a supposedly watertight category, especially where clusters are not tightly defined in multivariate space (Openshaw, 1995). The fuzzy geodemographics approach proposed by Openshaw (1989) is a potential resolution to this type of uncertainty. Slingsby et al. (2011) provide an example of this approach by visualising the propensity of each 2001 OAC Supergroup to be present in each OA across the UK. More widely within GIS, the term 'uncertainty', is used to denote that almost any representation is by its nature inherently incomplete (Longley et al., 2010). These problems are compounded when GIS representations seek to accommodate change over time (Plewe, 2002). In the context of geodemographics, the creation of temporal uncertainty indicators is an attempt to accommodate these problems. This allows fast

changing neighbourhoods to be identified whilst retaining the small level granularity of conventional geodemographic classifications for the others.

5.3. Population Change since 2001

The 2011 UK Census showed that the population of the UK was 63,182,178, a 6.9% increase from 2001 (ONS, 2012l). This increase in population can be broadly summarised as a gradual rise year-on-year, as indicated by Figure 5.1. Consistent with the classic filtering theory of urban geography (Hoyt, 1939), Sleight (2004) has suggested that this population change is of little over-all consequence for a geodemographic classification. Empirical studies (see Longley et al., 2011) suggest the longer term stability of regions, with most areas continuing to house the same social groups over time, even if the identities of the individuals themselves change. A temporal uncertainty indicator for a geodemographic classification however relies on the assumption that rates of change are not spatially consistent. Therefore quantification of the nature and degree of spatial change allows the validity of this assumption to be assessed.

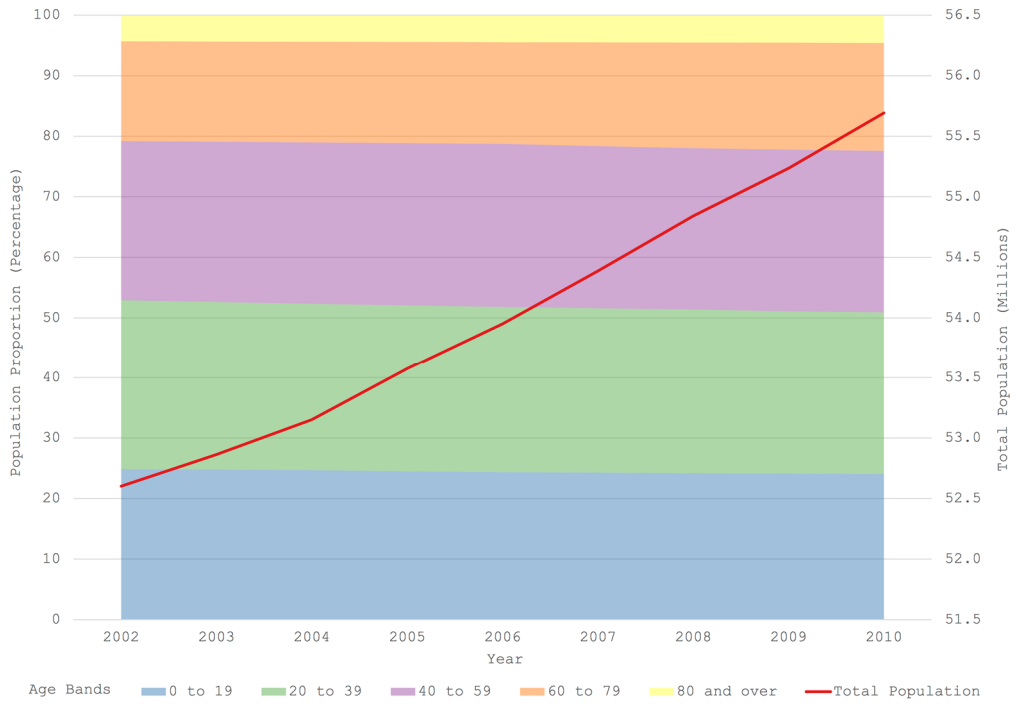


Figure 5.1: Population Change in England and Wales between 2002 and 2010
Source: ONS (2013d)

The defining feature of many area characteristics may be their continued stability, but this will not be true for every location. MYEs can be used to identify areas that have undergone change in real terms on a yearly basis. Despite not being able to quantify the change, MYEs at least allow for identification of small areas which have experienced fluctuations in their total population. Although equating changes in population size with geodemographic change is potentially flawed, it does seem likely that areas with the most unstable population counts will also be unstable from a geodemographics perspective. MYEs are therefore sufficient to highlight areas where the socio-demographic characteristics of an area, and therefore their geodemographic assignment, *may* have altered over time.

Population change is a spatial phenomenon. The 6.9% increase in UK population between 2001 and 2011 masks variation at smaller spatial scales. The variations seen across these levels provide an insight into how change differs between areas. For example in the recent UK inter-Census period the populations of England and Wales grew 7.1% (ONS, 2012a), Northern Ireland 7% (NISRA, 2012) and Scotland 5% (NRS, 2013c). Within this there were variations between the nine English regions and the other countries in the UK. Figure 5.2 uses MYEs from 2002 to 2010 to show how population change varied after 2001 between each region/country. London experienced the most change, with the North East of England showing the smallest increase over the decade. In real terms, the population of London increased by over 700,000, while that of the North East increased by less than 50,000. The uneven distribution of population change across England, and the rest of the UK, means that the impact on the socio-demographic characteristics of areas varied. As such, citing statistics from coarser spatial levels only illustrates general patterns, and data are required at finer levels of granularity to provide a more accurate understanding of the spatial variance of population change.

The availability of MYEs at the OA level in England and Wales means that changes in population for these countries can be observed down to the finest spatial scale. Figure 5.3 illustrates the maximum absolute deviation from the 2001 population for England and Wales between 2002 and 2010 for each OA. The bimodal distribution becomes more pronounced over time, and by 2010 every OA in England and Wales had experienced at least 1% population change, with 55.1% of OAs increasing in population. In London, 64.9% of OAs had an increase in population over this same time period, equating to a population increase of 9.1% (ONS, 2011). This increase was not consistent between boroughs, with the populations of Westminster and Tower Hamlets seeing increases of

24.8% and 18.3% respectively, while that of Brent decreased by 4.8%. Figure 5.4 shows how the increases and decreases in population varied by OA within each borough (see Figure 5.12 for borough names). The use of OAs allows for these small-scale changes at the neighbourhood level to be identified, although the significance this change had on the social, economic and demographic characteristics of each area varied depending on local conditions.

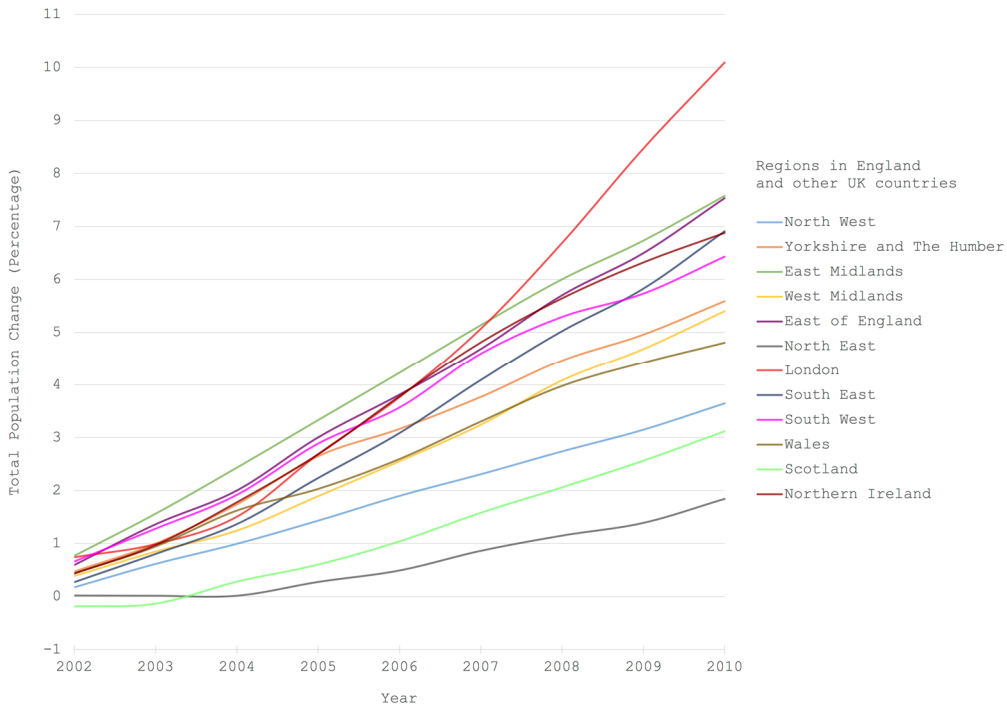


Figure 5.2: Population Change in England Regions and other UK countries between 2002 and 2010

Source: ONS; NRS and NISRA, Mid-Year Population Estimates

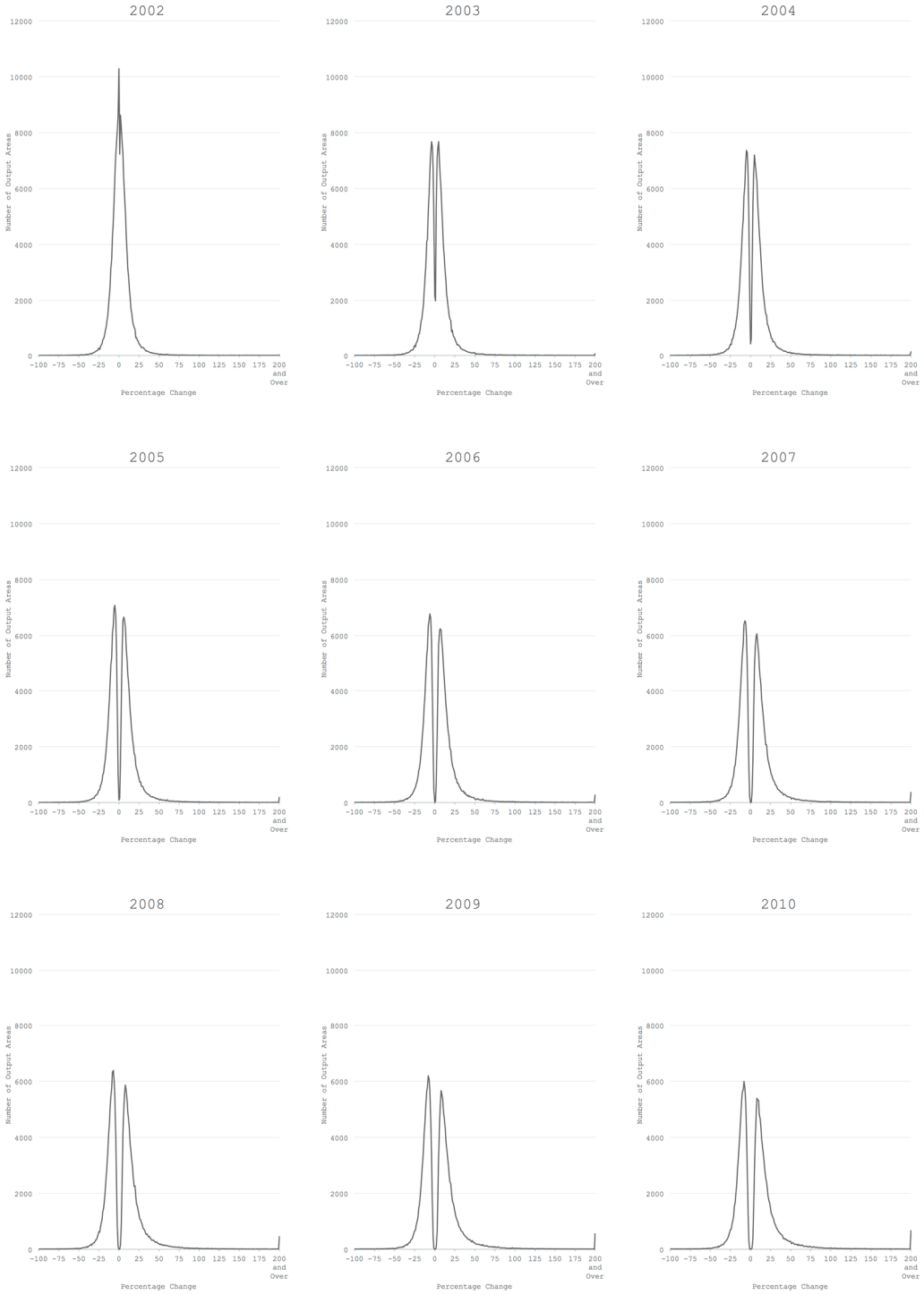


Figure 5.3: Maximum population change since 2001 in England and Wales from 2002 to 2010

Source: ONS, Mid-Year Population Estimates

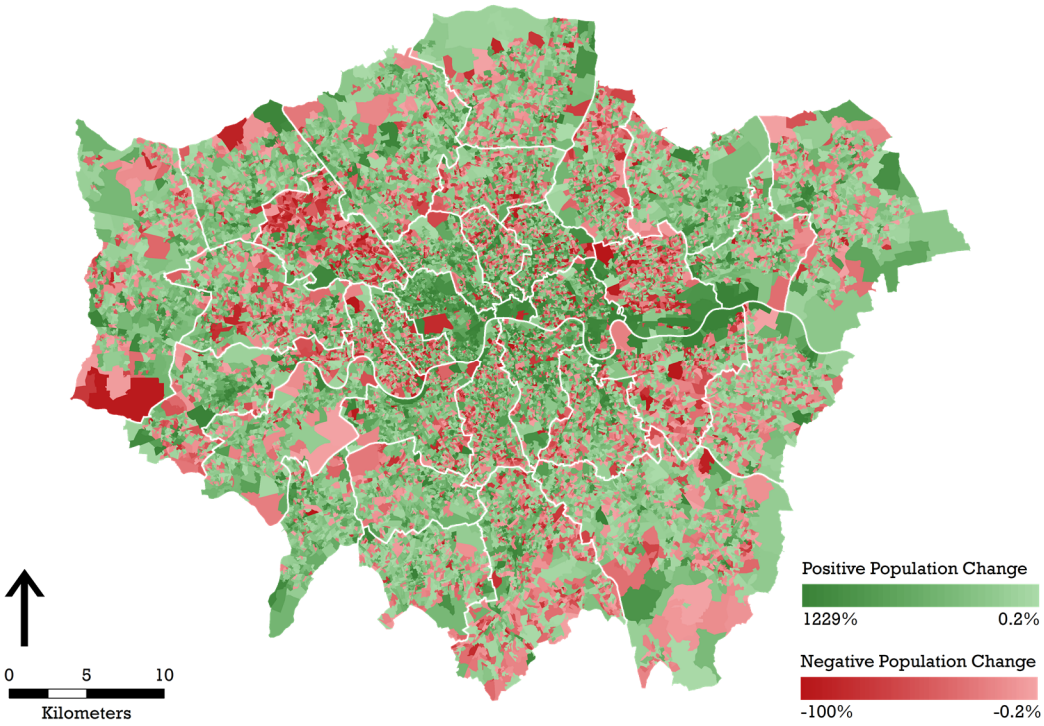


Figure 5.4: Population change in London between 2001 and 2010 by Output Area

5.4. Dwelling Stock Change since 2001

Population counts are not the only indicator of the changing characteristics of an area. The change in the physical environment, such as the number and type of dwellings, is another important factor. A dwelling can be defined as comprising a single household space or several household spaces sharing some facilities, and change in the dwelling stock enumerates the changes in any given area. There is likely to be a strong inter-relationship between changes in population and dwelling stock: for example, dilapidated housing stock might be cleared and replaced with new developments at different residential densities. In other cases existing housing stock may become occupied at higher residential densities by incomers, or redevelopment may not lead to changes in residential density. Changes in the total number of dwellings in an area can be the result of factors such as: new builds; demolition; conversions from houses to flats (or vice versa) and changes to and from residential use. Although unspecified in the statistics, the interaction between these factors led to a 7.2% rise to 22.8 million total dwellings across England between 2001 and 2010 (Department for Communities and Local Government, 2011b).

Figure 5.5 shows the increase in the number of dwellings across all English regions during the last inter-Census period, albeit to different extents. The North East, in conjunction with its population, witnessed the smallest increase. Unlike the changes in population, London had only a slightly larger than average increase in the number dwellings compared to the national average. Conversely, the South West and the East Midlands were the two regions that had large increases in the number of dwellings. The lack of correlation between the population and dwelling increases across England indicates differences in the dynamics driving these changes. For example, based on the statistics alone, it would appear that the growth of London's population outpaced the creation of new dwellings with the opposite appearing to be true for the South West. The total change in the number of dwellings provides an alternative view on the stability of an area when compared with population fluctuation alone.

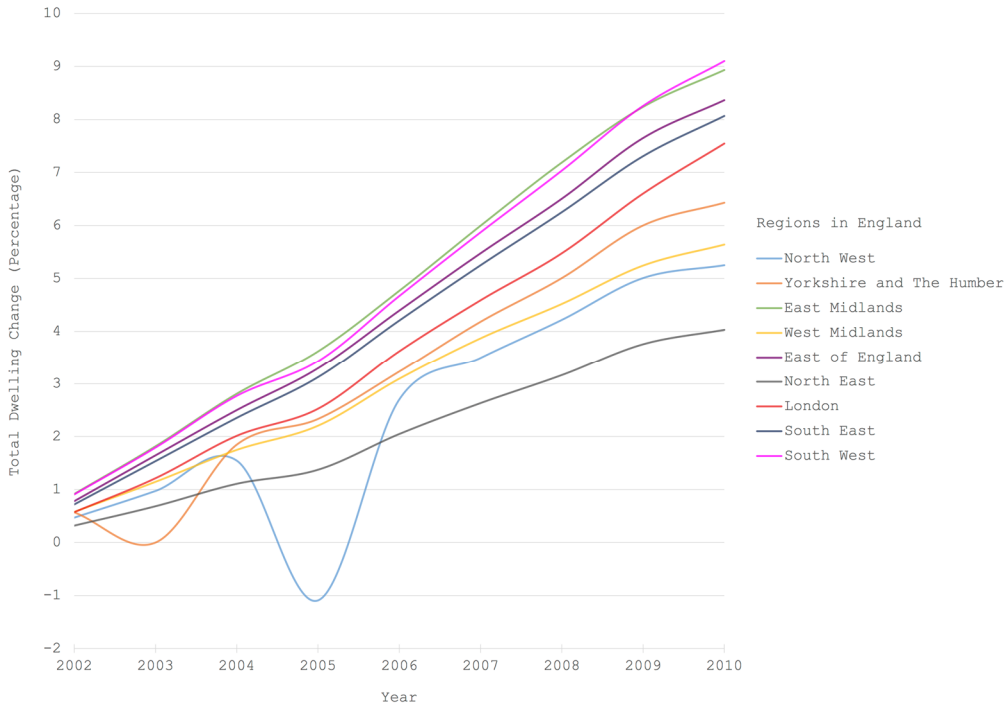


Figure 5.5: Dwelling change in English Regions between 2002 and 2010

Source: Valuation Office Agency, Dwelling Stock by Council Tax Band

The change in dwelling stock count can be further subdivided according to Council Tax band assignments. Council Tax bands are based on the capital value of residential property in England, Wales and Scotland. The Valuation Office Agency (VOA) assigns every residential property to a valuation band. In England for example these bands range from 'A' to 'H', where 'A' represents the cheapest dwellings and 'H' the most expensive (see VOA, 2008). Identification of the Council Tax bands of dwellings in an area is another means of further differentiating change and also a proxy for household asset holdings (Harper and Mayhew, 2012). Council Tax band assignments are released on a yearly basis in the public domain in England and Wales, although a property's band is only reassessed if it has been sold in the preceding 12 months (VOA, 2008). In England the assigned band is based on what the VOA estimates the property would have been worth on the 1st April 1991, even if the property was built after this date. The same is also true in Wales, although properties are revalued by reference to values at 1st April 2003 (Welsh Assembly Government, 2004). No equivalent data for Scotland or Northern Ireland is in the public domain.

Figure 5.6 shows the increase in the number of dwellings between 2001 and 2010 by Council Tax band for the regions in England. Although the North East had the smallest increase in the overall number of dwelling during this time period, it actually had the greatest increase in Band 'B', 'D', 'E', 'F' and 'G' dwellings. This was however coupled with a decrease of around 12,790 dwellings in Band 'A'. Although the North East has experienced a modest increase in the total number of dwellings, there was a shift away from the more affordable to the increasingly expensive type of dwelling. This differed from the South West where the largest increase was in affordable Band 'A' dwellings. Regions such as the South West and London tended to show relatively consistent increases across all Council Tax bands, while the North East, West Midlands and Yorkshire and The Humber had more variation. The changes these regions experienced were therefore not uniform, with modifications to the socio-demographic characteristics of these regions a reflection of these different dynamics.

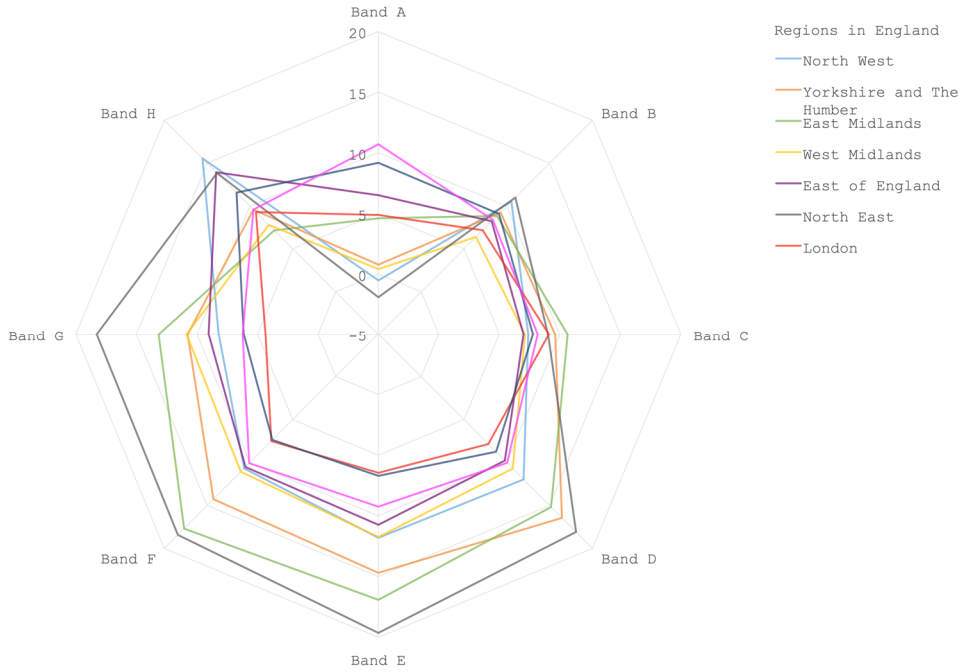


Figure 5.6: Percentage change of dwellings by Council tax band between 2001 and 2010 by Regions in England

Source: Valuation Office Agency, Dwelling Stock by Council Tax Band

The spatial variations of total population and dwelling stock change across the regions in England mean the interactions between these two important indicators vary from region to region. They provide independent views of the extent of change in any particular area. Figure 5.7 draws these two indicators together to assess how their combined change between 2001 and 2010 differs between the OAs in each English region. The values are calculated by dividing the OAs in England into three categories of change for both indicators. The bottom third which experienced the lowest amounts of change are classified as 'Low', the middle third 'Medium' and the top third 'High'. This creates nine different combinations of change, and the proportion of each of these combinations in the regions of England can then be assessed relatively. A similar pattern of change exists for most regions in England, although the notable exceptions to this are the North East and to a lesser extent the South West. In London a higher percentage of OAs experienced high population change, with smaller increases in dwelling stock to accompany this than the rest of England. In the North East, a total of 38% of OAs

experienced low dwelling and low or medium population change, a much higher proportion than the rest of England. London and the South West are notable for having a smaller proportion of OAs that experienced low dwelling and population change than the rest of England.

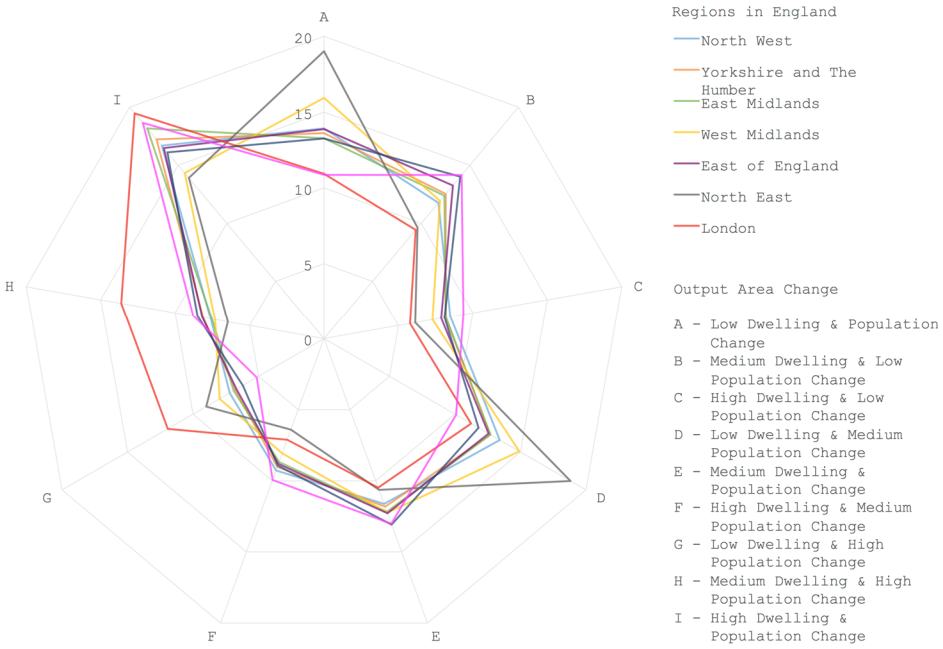


Figure 5.7: Population and Dwelling change between 2001 and 2010 by Regions in England

Source: ONS, Mid-Year Population Estimates and

Valuation Office Agency, Dwelling Stock by Council Tax Band

It can be surmised that the rate and type of change across the UK since 2001 is geographically inconsistent. At various spatial scales population and dwelling stock fluctuations are interacting with each other to create unique change environments. These different types of change, and the relative magnitude to which they occurred, cannot all be dismissed as irrelevant from a geodemographics perspective. The relative significance of the change identified, in determining the stability or otherwise of particular areas' socio-demographic characteristics, cannot be understood by looking at the statistics alone. Although this limits the use of MYEs and dwelling stock change to only identifying the fact that change has occurred, this is still significant in identifying areas that are likely to not conform to the assertion by Longley et al. (2011) that characteristics of most neighbourhoods do not change rapidly. Although this conclusion is based on relative, rather than absolute, values to distinguish how change differs between regions, it is still appropriate to suggest that MYEs and dwelling stock counts are appropriate to use as temporal uncertainty indicators for a geodemographic classification. The division of change into 'High', 'Medium' and 'Low' aids comprehension of the data, but masks the total range of variation. The population change in England between 2001 and 2010 ranges between 0% and 3,185% per OA, with the mean being 14.6%; while dwelling stock ranges between 0% and 23,400% with the mean being 12.5%.

5.5. Temporal and Spatial Uncertainty

The limited range of data sources available at the finest spatial scale makes monitoring change at the neighbourhood level challenging. MYEs and dwelling stock counts are rare examples of data sources that are made available at the OA level, albeit not with full UK coverage. It is this lack of data at both the finest spatial scale *and* with full geographical coverage that prohibits the creation of comprehensive temporal uncertainty indicators. Any attempt to counteract this would require additional techniques that seek to quantify other aspects of change that impact upon areas. Traditional measures of small area estimation (see Rao, 2005), such as regression models (Fay and Herriot, 1979), Bayesian methods (Congdon, 2010) or M-quantile models (Chambers and Tzavidis, 2006; Tzavidis et al., 2010) could offer synthetic estimates of change. The benefits of using such measures, however, are still reliant on multiple datasets being made available at the finest spatial scale and ensuring updates occur on a regular basis. The greater availability of data sources made available at the Lower Layer Super Output Area (LSOA) and equivalent levels present one possibility, but the less granular nature of these units' can

result in the obscuring of changes at the finest spatial level. The lack of full UK-geographical coverage is another issue. Not all statistics independently released by each UK country can be tabulated into a single dataset. An example of this is the Index of Multiple Deprivation (IMD), where the application of different methodologies by each country means that the data cannot be directly compared or combined to form a UK measure. The lack of data sources that fit these required criteria means that these estimation techniques cannot be used.

Utilising the limited number of data sources that are currently available at OA level to construct temporal uncertainty indicators provides a simple method of understanding the enduring relevance of a classification like the 2001 OAC. The benefit of not modifying the initial group assignments of the classification is that it maintains stability over its lifetime. Information provided on the likelihood of groups providing an accurate representation of an area is an addition to the classification, not instead of it. Although this method does not have the ability to reassign areas flagged as uncertain, it does offer an alternative to the ways in which commercial geodemographic classifications manage temporal change. The availability of Open Data, either made available or modelled at the finest level of granularity, means that this situation may change in the future. The outputs of the Beyond 2011 programmes may result in additional datasets being made available at the finest spatial levels, such as income, economic status and health status, but this would be dependent on the ONS obtaining access to relevant data sources (A. Calder, personal communication, 11th November 2013). The extent to which any changes in the availability of Open Data impact the ability for current and future geodemographic classifications to either be updated or construct more comprehensive temporal uncertainty indicators is unclear. Therefore the construction of temporal uncertainty indicators relies on utilisation of the best resources currently available.

MYEs and dwelling stock counts allow for the assessment of two key factors which impact upon the temporal reliability of a geodemographic classification assignment: (a) the extent to which the resident population size is likely to have changed, and (b) the nature and amount of recorded changes to the dwelling stock. In each of these cases changes in either or both of these indicators is likely to lead to differences in the demographic characteristics of OAs, along with changes in the numbers of individuals likely to bear these characteristics. While the totality of demographic change is unlikely to be captured by these two measures, nevertheless they provide a measure of the

reliability of local level demographic estimates and provide an insight into the enduring relevance of a geodemographic classification.

The production of separate MYEs on an annual basis for England and Wales, Scotland and Northern Ireland creates a disparity in the granularity at which the data are available. In England and Wales they are produced at OA level by single age band for both males and females. This is also true of Scotland, except that they are produced at Data Zone (DZ) level. Northern Ireland's output is different in that estimates are produced at Super Output Area (SOA) level for four age bands. While coverage of the UK is at varying levels of geography, the different measures remain useful because in each case they correspond to fine granularity Census geography. Dwelling stock data, classified by Council Tax band, is made available at the OA level only for England and Wales. The lack of freely available equivalent data for Scotland and Northern Ireland creates a disparity in UK coverage. Therefore utilisation of this data source to form a single temporal uncertainty indicator for the 2001 OAC results in the 21% of 2001 OAs assigned to Scotland and Northern Ireland being unrepresented.

The current geographic limitations of using dwelling stock data does not necessarily prevent it from being used solely or to form part of a temporal uncertainty indicator for England and Wales. However the disaggregation of total dwelling stock estimates into Council Tax bands to act as surrogate for housing wealth is hindered by a change in 2005, when new valuation bands and complete revaluation of all 1.3 million home in Wales was undertaken (Welsh Assembly Government, 2004), rendering them incomparable with their English equivalents. Nevertheless, while the band intervals are not compatible between the two countries they do allow for general changes in the values of property at the small area level to be seen; for example redevelopment and attendant upgrading of low-cost housing since 2001.

The limited number, and geographic inconsistences, of appropriate data sources means that any temporal uncertainty indicator created in the present data environment is a compromise. A single comprehensive temporal uncertainty indicator for the UK at the OA level is not possible. As such, separate indicators are required, with annual MYEs providing full UK coverage, and annual dwelling stock figures being an additional indicator for England and Wales only. Complications in the incorporation of Council Tax data for England and Wales, due to further subdivision of the UK following the Welsh revaluation, meant that this potentially rich data source could not be used. Although

outside the scope of this initial foray into the creation of temporal uncertainty indicators, Council Tax bands could prove to be useful in the future for creating country specific alternatives.

For each indicator, the maximum absolute deviation from the 2001 value over the period 2001 to 2010 was recorded. This allows the occurrence of change in circumstances in which an initial increase (or decrease) is subsequently compensated for by an ensuing decrease (or increase) to a value that might suggest that little change had occurred over the entire period to be recorded. As MYEs are not available for OAs outside England and Wales, using the coarser DZs for Scotland and SOAs for Northern Ireland was the best alternative. In addition to the utilisation of the population and dwelling stock change to form temporal uncertainty indicators, a combination of the two was created for England and Wales. The respective change shown by each indicator was standardised using z-scores and brought together to form an overall composite score, where data for each indicator are available (see Table 5.1).

5.6. Uncertainty and the 2001 OAC

The respective merits of using one of the three indicators created to identify temporal uncertainty with the 2001 OAC differs between areas. Although any direct comparison is limited to England and Wales only, the impact the differing dynamics of change in MYEs and dwelling stock counts have on different areas can nonetheless be quantified. The spatial variation in the relationship between MYEs and dwelling stock counts between regions impacts the amount of change identified by composite temporal uncertainty indicator.

Table 5.2 presents a confusion matrix of the areas of change as indicated by MYEs and dwelling stock counts. The data for England and Wales have been ranked and divided into deciles for each temporal uncertainty indicator. The OAs that share the same decile for both of the indicators suggest the uncertainty created in these areas derives equally from population and dwelling stock change. In the OAs where the temporal uncertainty indicator deciles do not match, it suggests that either population or dwelling stock change is driving the uncertainty in those areas, but not both. Decile 1 contains the OAs that have experienced the most change from 2001 to 2010 and decile 10 the least.

Table 5.1: Temporal uncertainty indicators

Temporal Uncertainty Indicator	Description	Geographical coverage	Data source
Population	Mid-year population estimates from 2002 to 2010 are used to calculate the maximum absolute deviation from the 2001 UK Census figures. This provides each OA a figure of maximum percentage change in the 2002 to 2010 period.	UK	Yearly mid-year population estimates provided by the ONS for England and Wales, NRS for Scotland and NISRA for Northern Ireland
Dwelling Stock	Dwelling stock counts from 2002 to 2010 are used to calculate the maximum absolute deviation from the 2001 figures. This provides each OA a figure of maximum percentage change in the 2002 to 2010 period.	England and Wales	Yearly dwelling stock by Council Tax band counts provided by the Valuation Office Agency
Composite	The population and dwelling stock percentage change are each standardised using z-scores. These figures are then added together to form a composite score for each OA.	England and Wales	Yearly mid-year population estimates and yearly dwelling stock by Council Tax band counts

Table 5.2: Population and Dwelling Stock temporal uncertainty indicators confusion matrix

		Population temporal uncertainty Indicator (Percentages)										
	Decile Number	**Decile 1**	**Decile 2**	**Decile 3**	**Decile 4**	**Decile 5**	**Decile 6**	**Decile 7**	**Decile 8**	**Decile 9**	**Decile 10**	**Total**
Dwelling Stock temporal uncertainty Indicator (Percentages)	**Decile 1**	4.36	1.14	0.80	0.66	0.63	0.54	0.53	0.51	0.47	0.45	10.10
	Decile 2	1.80	1.65	1.17	0.97	0.86	0.82	0.76	0.70	0.66	0.63	10.01
	Decile 3	0.95	1.41	1.25	1.11	1.07	0.97	0.88	0.85	0.79	0.73	10.00
	Decile 4	0.72	1.20	1.19	1.11	1.05	1.03	0.96	0.95	0.91	0.88	10.00
	Decile 5	0.55	0.95	1.10	1.11	1.07	1.09	1.05	1.04	1.03	1.01	9.98
	Decile 6	0.45	0.90	1.04	1.07	1.05	1.09	1.14	1.11	1.16	1.08	10.09
	Decile 7	0.42	0.81	0.93	1.00	1.09	1.07	1.16	1.15	1.17	1.14	9.93
	Decile 8	0.38	0.71	0.90	1.01	1.11	1.10	1.17	1.15	1.21	1.24	9.99
	Decile 9	0.47	1.23	1.66	1.94	2.09	2.28	2.39	2.53	2.56	2.74	19.89
	Decile 10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Total	10.11	10.00	10.04	9.96	10.01	9.99	10.05	9.99	9.96	9.90	100.00

Decile 10 for dwelling stock contains zero values as 11% of OAs in England and Wales experienced no change in dwelling stock over the past decade. The relationship between MYEs of population change and dwelling stock change shown in Table 5.2 lacks any strong correlation. One conclusion from this is areas that have experienced large population change are less likely to have experienced a large change in the total number of dwellings. This suggests that in areas where it is possible, using the composite temporal uncertainty indicator would be preferable. In part because areas which have experienced change in both population and dwelling stock have an increased likelihood of being misrepresented by their current geodemographic assignment.

The different aspects of change picked up by population and dwelling stock temporal uncertainty indicators, along with a combined composite measure provide an indication of geodemographic change in England and Wales over the past decade. Table 5.3 identifies the thresholds of change, beyond which the 2001 OAC is deemed unreliable. These threshold values, like many decisions in geodemographic classification, are subjective. The threshold values were based on the identification of areas that were within one standard deviation, around 68.2% of the UK's OAs, and classing them as unchanged. Manual intervention was required to decide upon the final threshold values to allow for the greatest compatibility between the temporal uncertainty indicators possible, but also limit the areas classified as uncertain to locations where the more extreme changes in local characteristics have taken place. Overall, the percentage of OAs classed as uncertain using each of the three temporal uncertainty indicators ranges from 21% to 29%. Figure 5.8 displays the distribution of change for each temporal uncertainty indicator, with the additional inclusion of population change for England and Wales, Scotland and Northern Ireland for reference.

Table 5.3: Threshold distribution of temporal uncertainty indicators

Temporal Uncertainty Indicator	Negative Threshold Value	Positive Threshold Value	Average percentage of OAs Below Threshold	Average percentage of OAs Above Threshold	Below Threshold to Above Threshold Ratio
Population	-15%	20%	71	29	2.5:1
Dwelling Stock	-15%	20%	74	26	2.8:1
Composite	-0.8	0.4	79	21	3.7:1

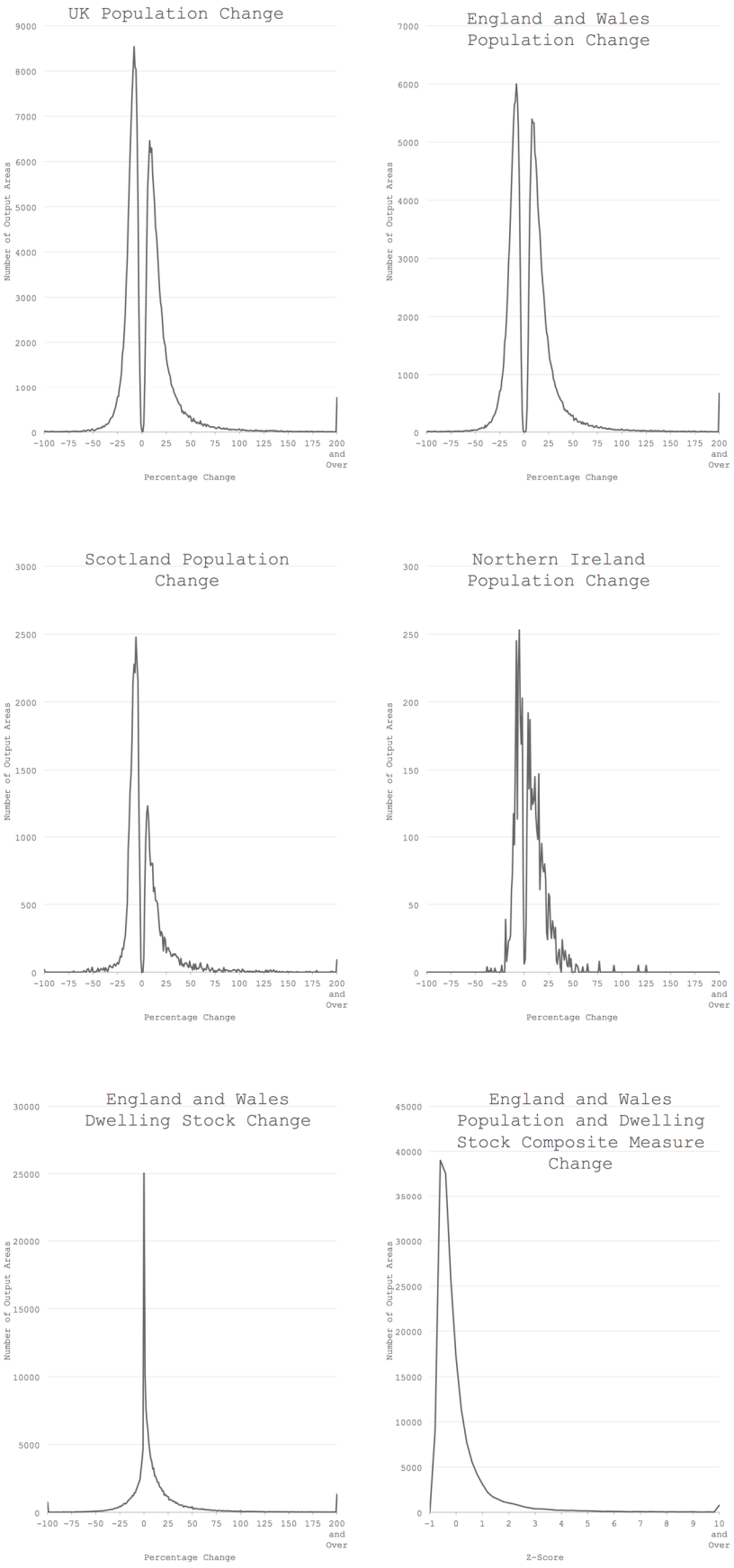


Figure 5.8: Change distribution in 2010 of temporal uncertainty indicators

Figures 5.9, 5.10 and 5.11 illustrate how the combinations of the three temporal uncertainty indicators can be used alongside threshold values. Each map is a density equalising cartogram (Gastner and Newman, 2004), in which the area of every OA has been rescaled in direct proportion to its total population in 2010. The 2001 OAC has been visualised to show how the change identified by the indicators varied between the assigned geodemographic groups.

The population temporal uncertainty indicator shows that a large number of OAs in the Greater London area experienced population change over the threshold value. The rest of England and Wales has a fairly even distribution of above threshold values, but other urban areas such as Manchester and Birmingham dominate their respective local areas. The dwelling stock temporal uncertainty indicator provides a different picture of change in England and Wales. OAs that have experienced change greater than the threshold values are predominately distributed in the South East and South West of England, with 34% of all OAs in these two regions experiencing dwelling stock change greater than the threshold value.

The composite temporal uncertainty indicator has a different geographical distribution again; although, as Table 5.3 indicates, it designates fewer OAs in total as uncertain when compared to the two other temporal uncertainty indicators. Of the population and dwelling stock temporal uncertainty indicators, it is the population measure that is geographically dispersed across England and Wales, albeit with higher concentrations of change in urban areas. Change in the dwelling stock indicator is particularly marked in the South East and South West of England. The composite indicator also suggests that the greatest incidence of change is in the South East and South West of England, and also with concentrations in urban areas across England and Wales.

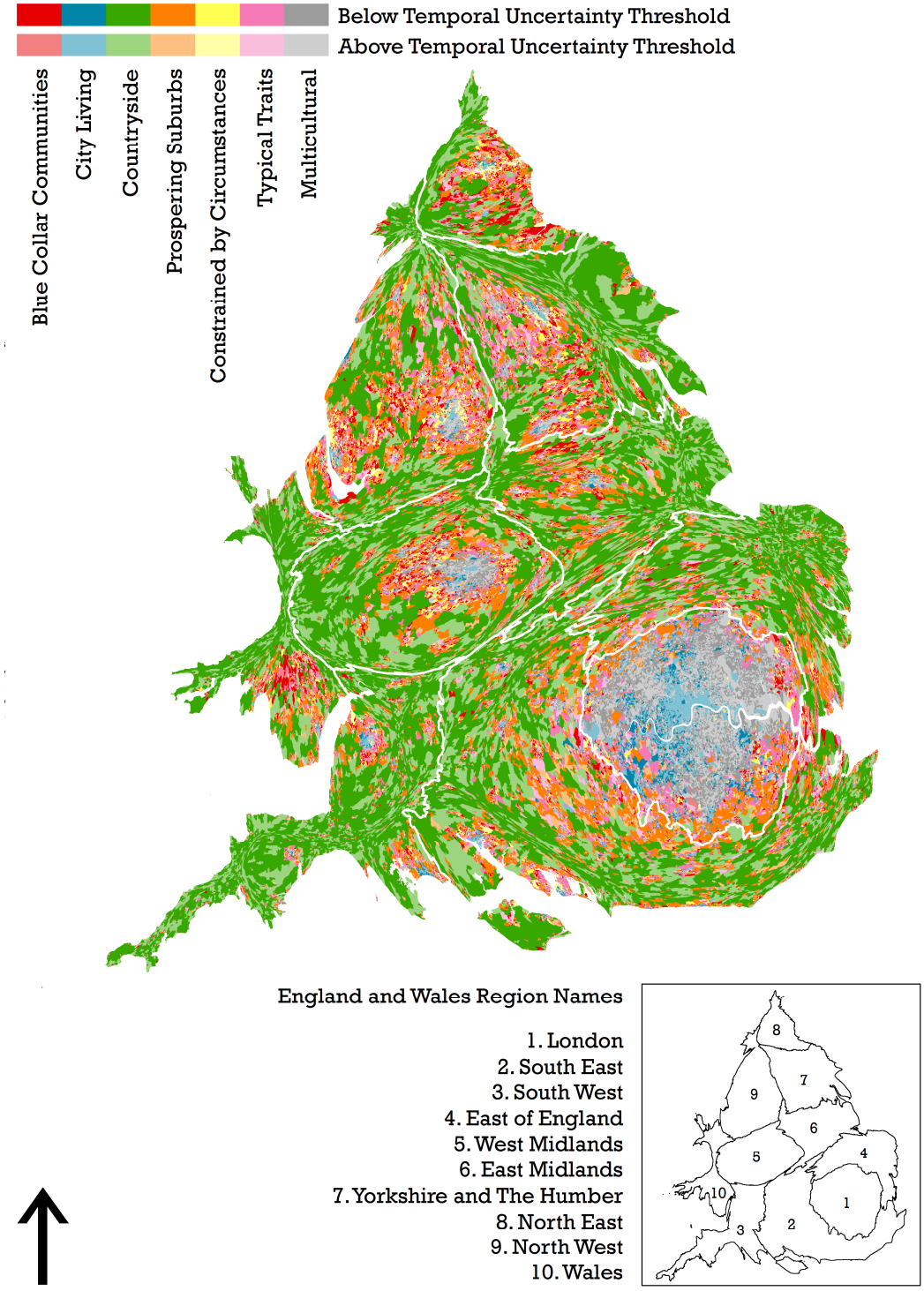


Figure 5.9: Threshold of Population Change temporal uncertainty indicator in England and Wales for the 2001 OAC viewed as a cartogram

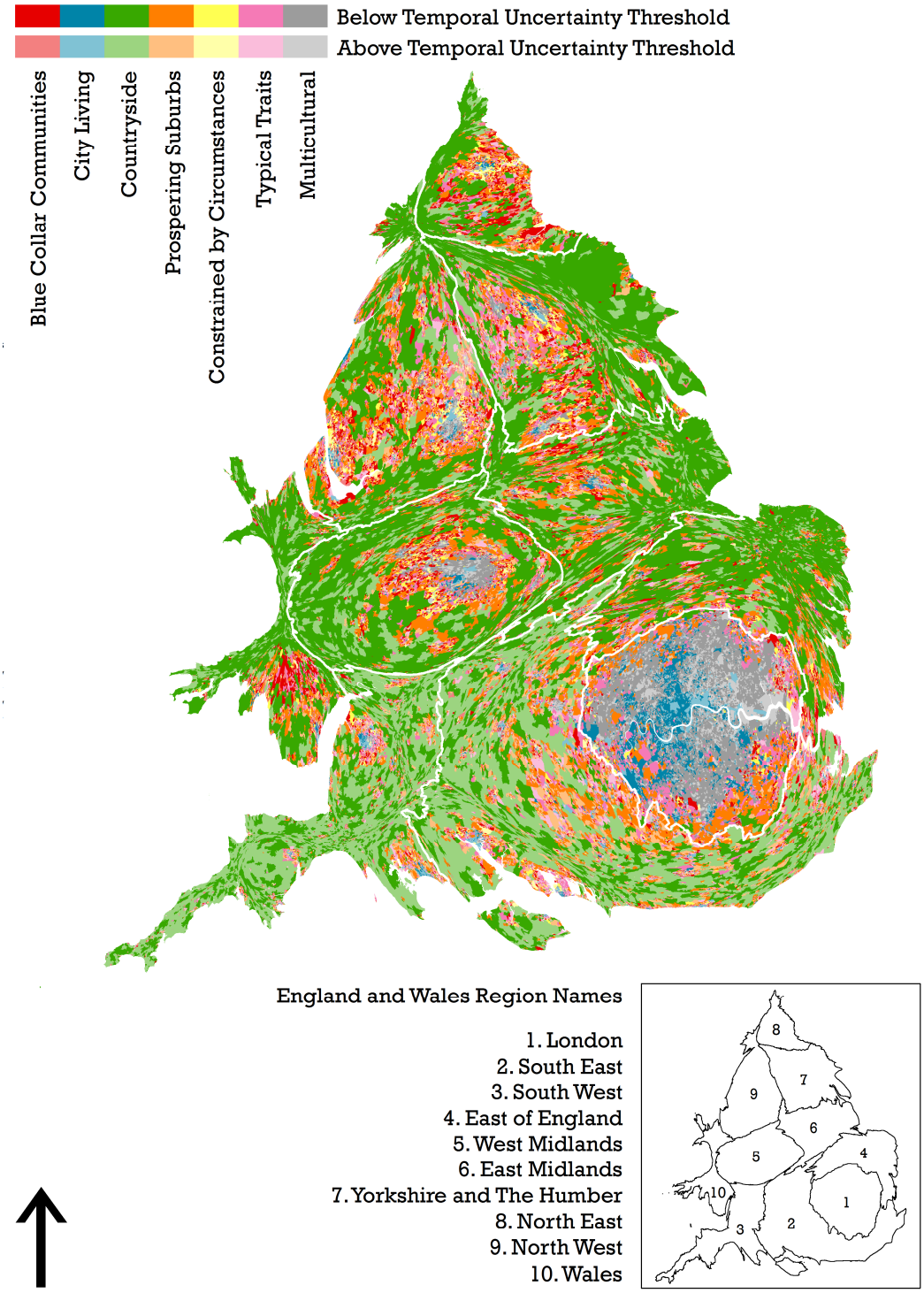


Figure 5.10: Threshold of Dwelling Stock Change temporal uncertainty indicator in England and Wales for the 2001 OAC viewed as a cartogram

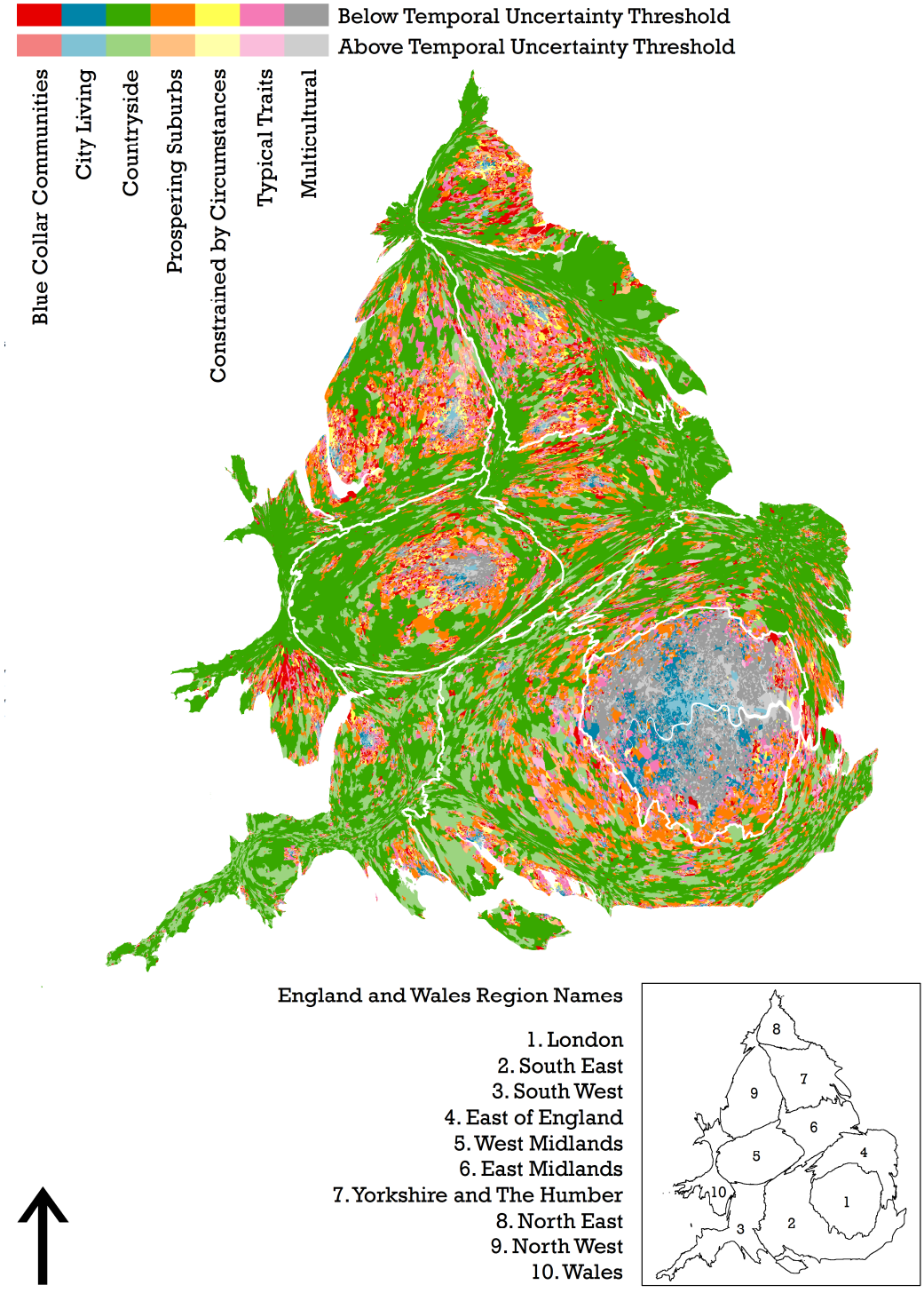


Figure 5.11: Threshold of Population and Dwelling Stock Composite Change temporal uncertainty indicator in England and Wales for the 2001 OAC, viewed as a cartogram

As shown in Figures 5.9, 5.10 and 5.11, the three temporal uncertainty indicators identify areas across England and Wales which are subject to the most change. Table 5.4 supplements this, tabulating these results by the 2001 OAC Supergroups. A third of the OAs identified as having experienced change above the population temporal uncertainty indicator threshold are in the 'Typical Traits' and 'Multicultural' Supergroups, suggesting that change in the incidence of these two Supergroups is heavily driven by population size change. This provides only part of the picture as 'Typical Traits' is also influenced by changes to dwelling stock, as over 20% of OAs identified as having above threshold change to dwelling stock are located in this Supergroup. Compared to the 10% figure for the 'Multicultural' and the 21% for the 'Prospering Suburbs' Supergroups, it is clear that different combinations of change drive the uncertainty of geodemographic types to varying extents. The composite temporal uncertainty indicator provides only a slight variation to the distributions seen with the dwelling stock measure. While arguably just an artefact of the threshold values used the indicator, it could be the case that the more extreme change, and therefore uncertainty, seen across the 2001 OAC Supergroups is driven more by dwelling stock change than just population change alone.

Table 5.4: Above threshold percentage distribution of the temporal uncertainty indicators by 2001 OAC Supergroup

Temporal Uncertainty Indicator / 2001 OAC Supergroup	Population temporal uncertainty Indicator	Dwelling Stock temporal uncertainty Indicator	Composite temporal uncertainty Indicator
Blue Collar Communities	11	15	15
City Living	13	7	7
Countryside	12	14	12
Prospering Suburbs	14	21	21
Constrained by Circumstances	15	11	11
Typical Traits	17	21	20
Multicultural	17	10	13
Total*	100	100	100

* Figures may not sum exactly due to rounding

Table 5.5 illustrates that across the regions of England and Wales there are geographic variations in the percentage of OAs that have greater than threshold levels of change for each temporal uncertainty indicator. London for example has 46% of OAs classed as uncertain if using the population temporal uncertainty indicator, but only 19% or 21% if using the dwelling stock or composite temporal uncertainty indicators respectively. A similar pattern can be found in Yorkshire and the Humber along with the East Midlands. Conversely, in the South East and South West of England the dwelling stock change measure expresses greater uncertainty than the other two indicators. The variability seen in uncertainty picked up by the population and dwelling stock temporal uncertainty indicators between the regions is not repeated for the composite measure. This is due to the combination of the population and dwelling stock temporal uncertainty indicators cancelling each other out. The population and dwelling stock temporal uncertainty indicators have a range of 19% and 18% respectively between the regions in the amount of uncertainty inferred. For the composite temporal uncertainty indicator this is just 5%, suggesting this measure has an increased stability across England and Wales.

Table 5.5: Above threshold percentage distribution of the temporal uncertainty indicators by regions in England and Wales

Temporal Uncertainty Indicator / Regions in England and Wales	Population temporal uncertainty Indicator	Dwelling Stock temporal uncertainty Indicator	Composite temporal uncertainty Indicator
East of England	29	29	21
East Midlands	30	21	19
London	46	19	21
North East England	27	25	22
North West England	30	29	24
South East England	29	32	20
South West England	30	37	22
Wales	29	29	21
West Midlands	29	23	19
Yorkshire and the Humber	29	19	21

Figure 5.13 presents the composite temporal uncertainty indicator for London (for the names of the Boroughs see Figure 5.12) in 2010, broken down by 2001 OAC Supergroups. 22% of London OAs classified as 'Multicultural' exceed the threshold, as do a similar percentage of 'City Living' neighbourhoods. The 'Multicultural' and 'City Living' Supergroups together comprise over 75% of London's OAs, and only 15% of OAs assigned to any of the other five Supergroups exceed the threshold. Change in the 'Multicultural' group is predominantly found in the east of London, in the Boroughs of Tower Hamlets, Newham, Hackney and Barking and Dagenham. The City of London and the City of Westminster in the centre of London are where the greatest change in the 'City Living' assignments are found. The distribution of change in the other five Supergroups shows no distinct pattern. No single area has a particularly high concentration of change, with the areas identified as uncertain being found in isolated pockets situated around the outer-Boroughs of London. These results reflect the dominance of the 'Multicultural' and 'City Living' Supergroups in London, with over three-quarters of the capital falling into one or other of these two clusters. The dominance of these two Supergroups means large geographic areas can be identified as being uncertain. Such areas do not exist for the other Supergroups due to their more sporadic geographical distribution across London.

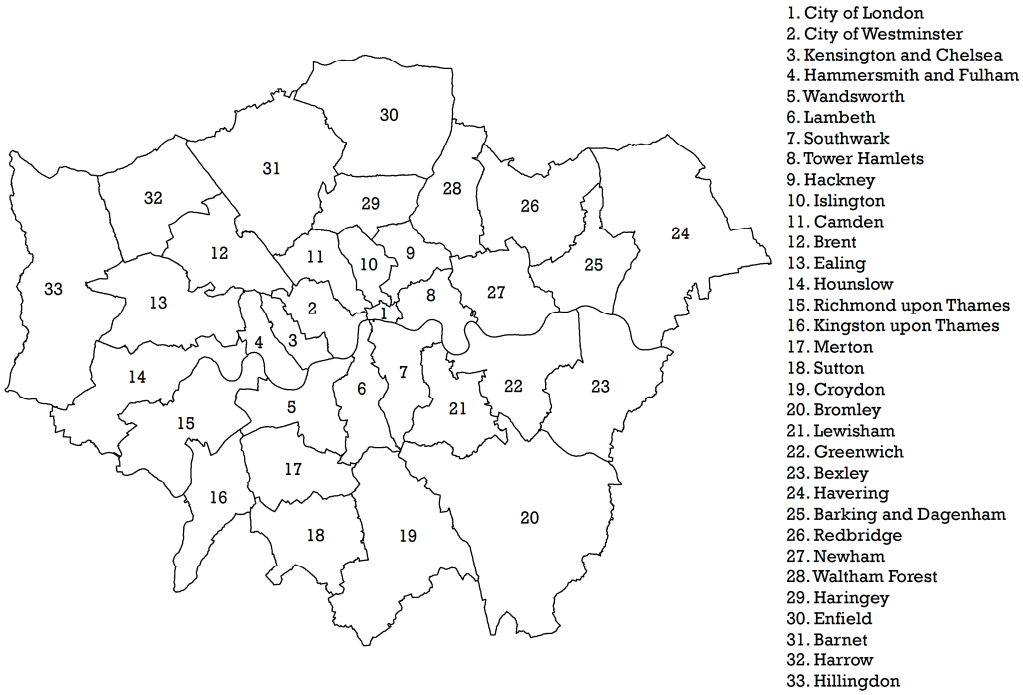


Figure 5.12: London Boroughs and the City of London

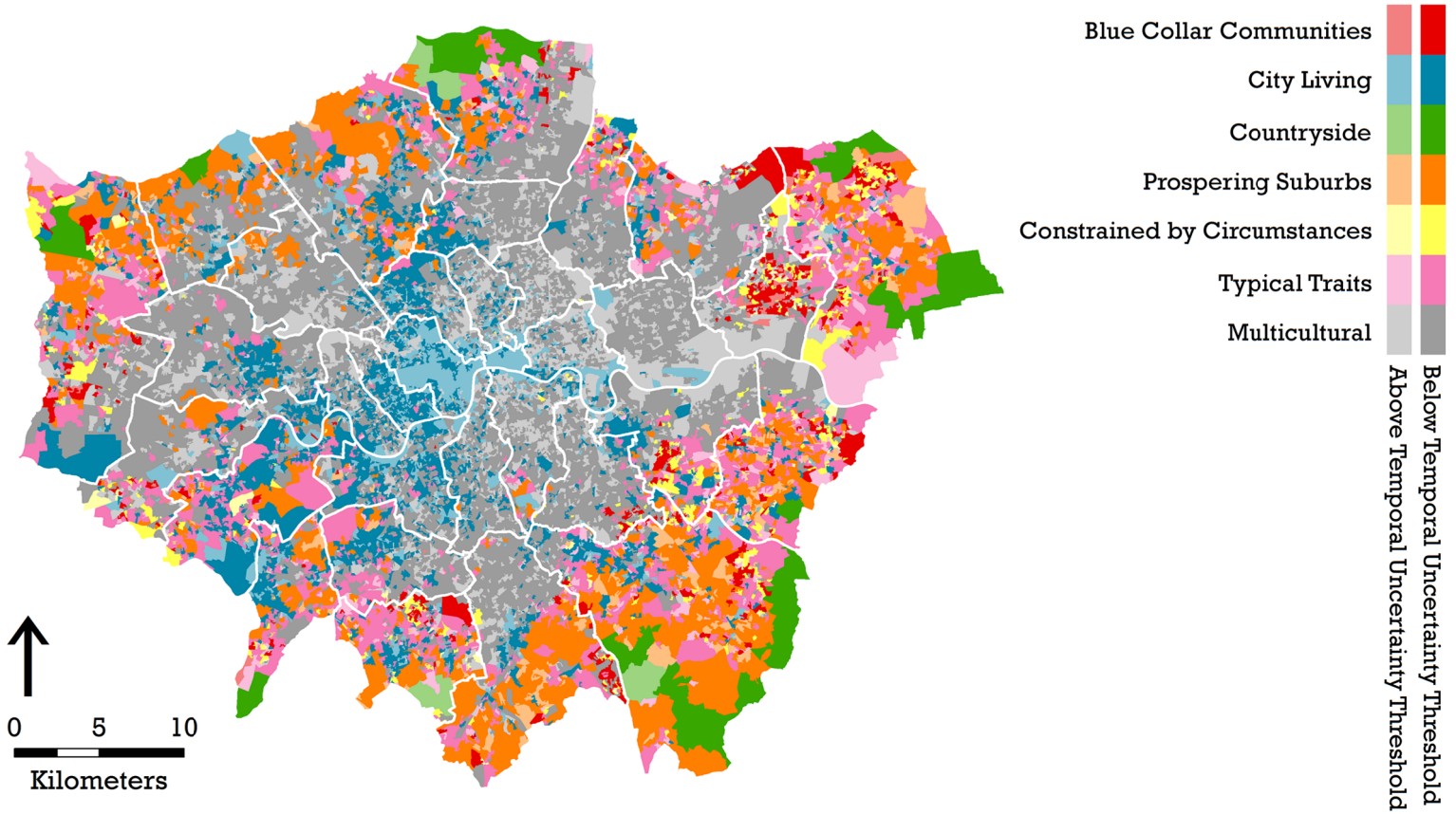


Figure 5.13: Distribution of 2001 OAC Supergroups OAs falling above and below the composite temporal uncertainty indicator threshold values in 2010

As the only temporal uncertainty indicator with coverage for Scotland and Northern Ireland is that derived from MYEs for population size, there is limited scope to critique the effectiveness of the indicator in these countries beyond analysing its findings, and comparing it to areas in England and Wales. Figure 5.14 highlights the example of Glasgow and its surrounding area for each of the seven Supergroups in the 2001 OAC. 'Constrained by Circumstances' is the dominant Supergroup in Glasgow, with 55% of OAs assigned to it. This dominance does not however translate to an increased propensity for change as only 31% of OAs assigned to this group are classed as uncertain. In addition, there is no distinct patterning to this change. It does not appear that consolidations of DZs exhibiting change characteristics of uncertainty have developed over the past decade, unlike, for example, the 'Multicultural' or 'City Living' groups in London. There is a similar pattern with the other Supergroups in Glasgow, where no distinct concentrations of change have developed. The exception to this is the 'City Living' group where the majority of uncertain areas are located in the centre of Glasgow. This apparent difference between Glasgow and London can in part be explained by the total proportion of OAs that have been classed as uncertain. London's dominant 'Multicultural' Supergroup has over half of the OAs assigned to that group classed as 'uncertain' when using the population temporal uncertainty indicator. In terms of overall population change, London is a more rapidly changing city than Glasgow with 46%, compared to 27% of OAs being above the uncertainty threshold.

Analysis of how the variance in uncertainty across the UK impacts upon the likely changes in distributions of the 2001 OAC Supergroups between 2002 and 2010 is shown in Figures 5.15 and 5.16. The limited change in 2002 increases steadily through to 2010, with the 'Blue Collar Communities' and 'Prospering Suburbs' Supergroups experiencing limited change relative to the 'Multicultural' and 'City Living' groups. It is evident that neighbourhoods assigned to different geodemographic groups have differing propensities to change. Table 5.6 presents the percentage changes for each of the 2001 OAC Supergroups between 2002 and 2010, broken down into the constituent parts of the UK. The population assigned to each of the 2001 OAC Supergroups has experienced greater change in England and Wales, with the change in Scotland and Northern Ireland being smaller in magnitude. These results suggest that OAs in England and Wales are proportionally more likely to have changed in the period since 2001. This over-all change can be further sub-divided according to 2001 OAC Supergroup to accommodate both the effects of location and geodemographic characteristics and to determine the level of uncertainty associated with use of the classification.

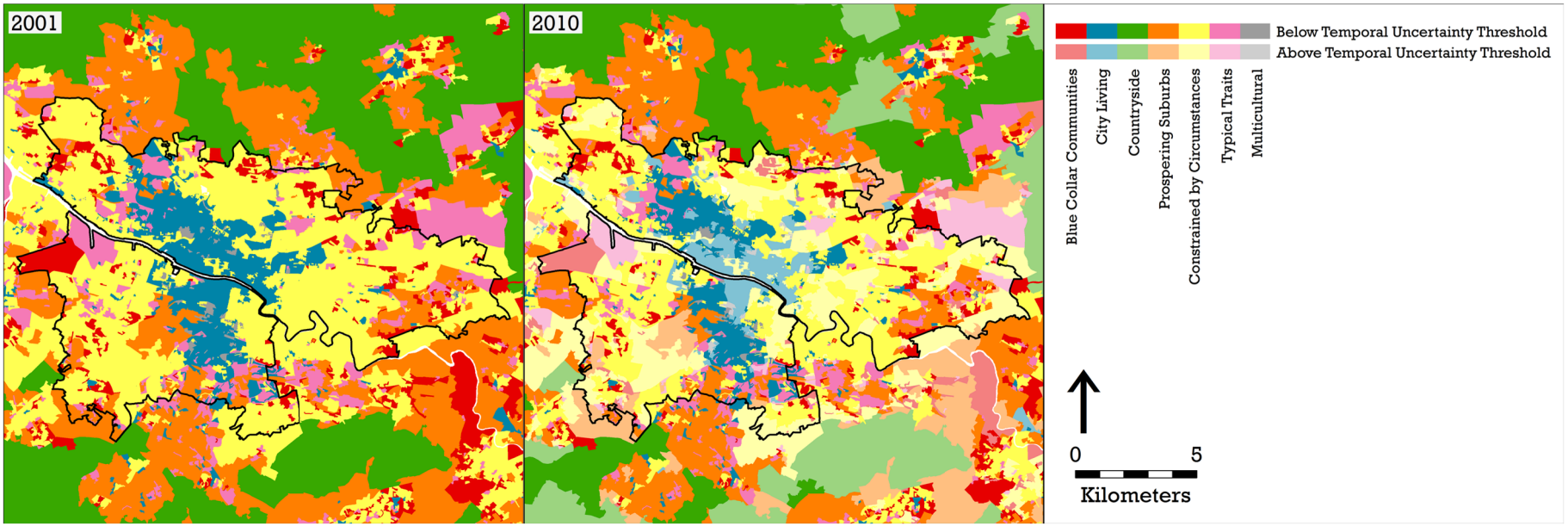


Figure 5.14: 2001 OAC Supergroups in the Greater Glasgow region falling above and below the population temporal uncertainty indicator threshold values

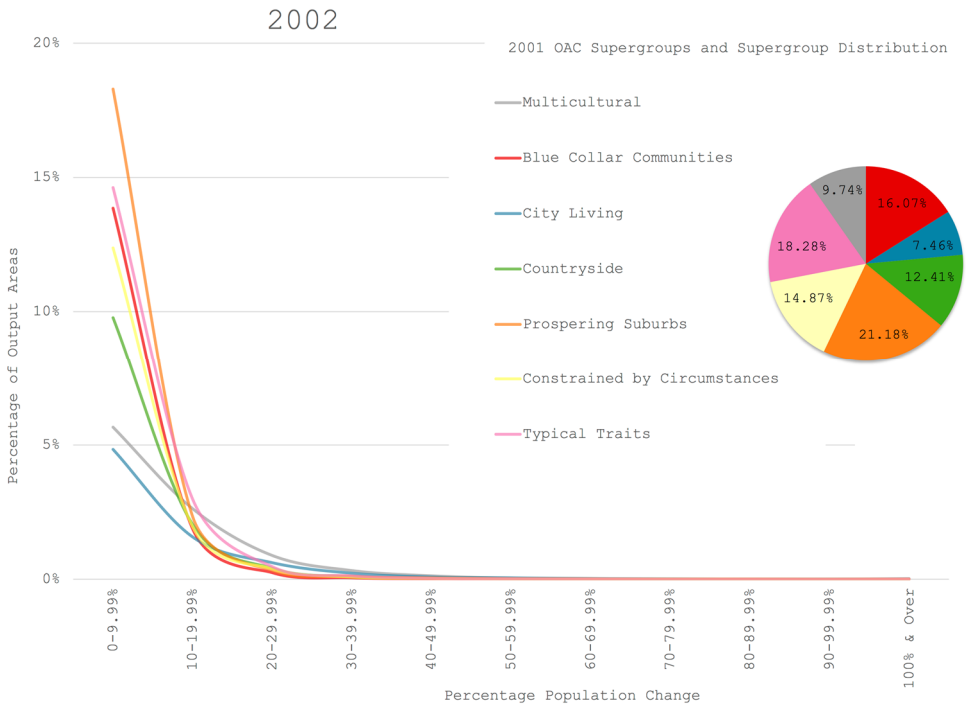


Figure 5.15: UK distribution of the population temporal uncertainty indicator by 2001 OAC Supergroups in 2002

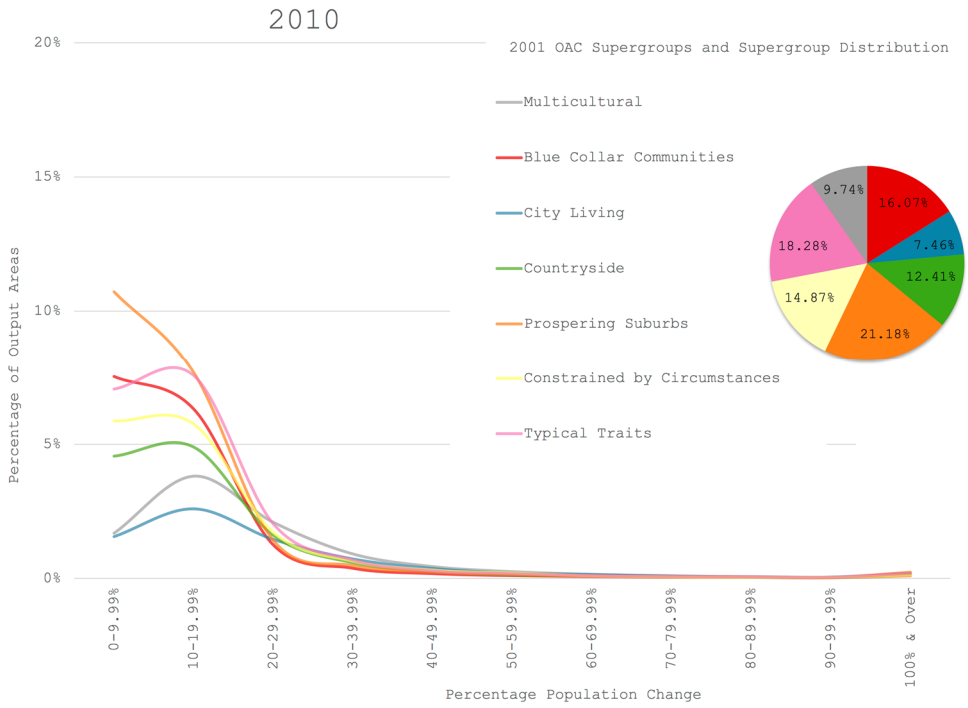


Figure 5.16: UK distribution of the population temporal uncertainty indicator by 2001 OAC Supergroups in 2010

Table 5.6: Percentage distribution change of the population temporal uncertainty indicator between 2002 and 2010 by 2001 OAC Supergroup for England and Wales (EW) and Scotland and Northern Ireland (SNI)

		2001 OAC Supergroup													
		Blue Collar Communities		**City Living**		**Countryside**		**Prospering Suburbs**		**Constrained by Circumstances**		**Typical Traits**		**Multicultural**	
		EW	SNI	EW	SNI	EW	SNI	EW	SNI	EW	SNI	EW	SNI	EW	SNI
Percentage Population Change	**0.00 to 9.99**	-47	-42	-78	-54	-54	-51	-43	-36	-59	-47	-54	-36	-70	-67
	10.00 to 19.99	173	3108	26	1173	105	2778	195	968	105	1208	138	1823	44	591
	20.00 to 29.99	341	4040	104	3742	251	2176	274	1788	241	1374	319	1400	135	1000
	30.00 to 39.99	479	100	167	100	411	4900	281	1344	409	100	424	100	179	100
	40.00 to 49.99	695	1300	281	100	402	6100	322	1150	507	12350	586	100	244	100
	50.00 to 59.99	785	100	246	100	423	100	427	1700	600	100	728	3300	436	100
	60.00 to 69.99	1671	100	419	100	470	100	413	100	1850	100	900	100	478	100
	70.00 to 79.99	2375	100	359	100	836	100	518	100	867	100	1656	100	848	0
	80.00 to 89.99	3650	100	494	100	644	100	389	100	3050	100	1683	100	636	100
	90.00 to 99.99	2850	100	457	100	1050	100	600	100	1900	100	2800	100	738	100
	100.00 and over	5900	100	929	100	1195	100	1381	100	9100	100	2271	100	1204	100

The results from using these temporal uncertainty indicators with the 2001 OAC need to be tempered with the qualification that there is a degree of uncertainty associated with the change data sources themselves. The reliability of using MYEs as a change indicator over time is influenced by the methodology used to produce them. There are slight differences in the methods used to calculate the MYEs in each UK country although the three responsible organisations each use a common cohort component method to update the population base (ONS, 2010b). Essentially the Census is used as a population base and then each year births are added and deaths subtracted based on data from the General Register Office. Internal migration estimates for the UK are based on three administrative datasets: National Health Service Central Register, Patient Register Data Service and Higher Education Statistics Agency (ONS, 2013d). In England, Wales and Scotland, data on international migration come from the International Passenger Survey (IPS), the Labour Force Survey (LFS) and Home Office data on asylum seekers and their dependants (ONS, 2010b). In Northern Ireland the inflows are estimated from the list of patients registered with a family doctor and the outflows from the number of people who have de-registered: data from the Irish Quarterly Household Survey (QNHS) are also used to estimate migration to the Republic of Ireland (NISRA, 2011). Further issues associated with migration statistics are discussed in ONS (2011) for England and Wales, in GROS (2010) for Scotland and in Dignan et al. (2010) and IJpelaar et al. (2011) for Northern Ireland.

The release of 2011 UK Census data provided an opportunity to critique the accuracy of MYEs created over the last inter-Census period. The ONS identified that for England and Wales the population of males aged 10 to 19 and 30 to 39 is larger than that suggested by the population estimates for March 2011, while the opposite is true for the male population aged 20 to 29 (ONS, 2012m). Amongst other discrepancies, the March 2011 population estimates are too high for the 25 to 29 age group in some university areas (ONS, 2012m). The accuracy of the MYEs is thus likely to be spatially variable, with knock on consequences for any temporal uncertainty indicator that utilises them. Dwelling stock counts, and their change since 2001 are at present the only viable alternative indicator to the uncertainty of the 2001 OAC. Unlike MYEs these are based on enumeration of all residential properties, rather than an estimate, so a greater level of certainty may be attached to the figures, although this does not make them more important in evaluating the broader picture of temporal change. Although issues no doubt persist with the use of these data sources, and in particular the potential inaccuracies of the MYEs, they still represent the best available sources for temporal

uncertainty indicators. As such, any negatives that may exist are outweighed by the potential benefits they can provide.

5.7. Conclusions

The population of the UK increased by 6.9% to 63,182,178 between 2001 and 2011; with the total number of dwellings in England and Wales also increasing between 2001 and 2010 by 7.2% to 24,188,815. This change is geographically variable across a range of scales, and the patterning of change is distinctive between the regions of the UK – so far as one can reasonably tell given the inherent vagaries in the source data. The use of the temporal uncertainty indicators with the 2001 OAC has shown that in the majority of locations, the original geodemographic assignment of the classification remained valid. The limited change to the population and dwelling stock dynamics across large areas of the UK during the last inter-Census period provides empirical evidence that temporal uncertainty indicators can be used to gauge the stability of geodemographic classifications over time. It can therefore be concluded that costly and time-consuming updating through ancillary sources used by commercial providers is only required for some areas of the UK. Despite this, many users believe that the most useful classifications require the most current data. The findings from the application of the temporal uncertainty indicators would appear to conform to Hoyt's (1939) notion of filtering in urban structure, whereby the social, economic and demographic structure of neighbourhoods remains stable over time, even if the identities of the residents themselves turn over much more rapidly. A practical implication of this is that users of the 2001 OAC can have increased confidence in the use of the classification in the majority (on average 74%) of areas where analysis suggests change has been more muted. Although, for areas where this is not the case, such as the London Boroughs of Tower Hamlets, Newham, Hackney and Barking and Dagenham it would indicate greater caution when using the 2001 OAC.

The successful creation of temporal uncertainty indicators for the 2001 OAC suggests that they should form part of a new geodemographic classifications that cannot employ the updating techniques used by commercial operators. In comparison to utilising a classification in isolation, the advantage of knowing which areas may no longer resemble their initial classification designation becomes clear. The user becomes aware of the need to investigate such areas using alternative data sources in order to better understand any change in an area's dynamics and make more informed interpretations.

Addressing some of the perceived inadequacies of academic geodemographics when compared to commercial alternatives by using subjectively defined thresholds to identify significant change to population, dwelling stock or a combination of the two, provides a credible alternative to the commercial offerings. This should help to reduce the perception that entire geodemographic classifications which are not updated using traditional methods become out dated quickly.

The use of temporal uncertainty indicators should however be tempered in the knowledge that they themselves are not infallible. MYEs in particular incorporate an element of uncertainty due to their methodological underpinnings. In addition, there is the possibility that the greater the estimated population change, the greater the associated uncertainty. Data issues in Scotland and Northern Ireland further compound these qualifications, where updating is only possible at higher levels of granularity. The problem of data-mismatch between countries in the UK, like that of the availability of more datasets at the finest spatial levels, is unlikely to be resolved in the near future. Aside from the issues relating to data availability and quality, the analysis and conclusions drawn from the use of the temporal uncertainty indicators with the 2001 OAC is itself uncertain. Assumptions made in linking population and dwelling stock data to a wider range of population characteristics are not necessarily correct. However, as the underpinning methodology is open and transparent there are clear benefits in utilising the methods discussed in this chapter compared to those used in the commercial sector.

Looking prospectively at the 2011 OAC, the use of temporal uncertainty indicators will be an important element in both the successful uptake and continued use of the classification. It can be envisaged that if more data sources are made available at both the finest spatial units and with wide UK geographic coverage, they could be incorporated into more comprehensive indicators.

Additional small area change measures might be developed from Open Data sources in the future, although research would be required into the volatility and reporting bias in small area estimates before usable small area measures were developed. Beyond this, and depending on the findings from the Beyond 2011 programme, there is a possibility that as increasing amounts of relevant Open Data become available, so improved methodologies may be devised in order to update classifications, and indeed identify the point at which an entire classification needs to be re-engineered. This would represent a

solution to the temporal and spatial instability of non-commercial geodemographic classifications, bringing them closer to the methods currently used by the commercial companies. However, as expressed by the analysis in this chapter, this is currently only a hypothetical possibility.

Chapter 6

Methodology for the 2011 Area Classification for Output Areas

6.1. Introduction

The purpose of this chapter is to outline the general procedures and the methodological underpinning used to create the 2011 Area Classification for Output Areas (2011 OAC). Section 6.2 explains the process of cluster analysis and how it relates to creating area classifications, with additional details relating specifically to the building of the 2011 OAC. Section 6.3 gives an overview of the methodology used to create the 2011 OAC, before Sections 6.4 to 6.8 provide specific details of the different steps taken to create the classification. Section 6.4 details the variable selection process and the rationale for why the 2011 UK Census was the sole data source used for the 2011 OAC. It also describes the Census data that were used to form the list of candidate variables that were identified for possible inclusion in the classification.

Section 6.5 explains the data preparation techniques used on the raw Census data; to use these data in a geodemographic classification they must be converted, transformed and standardised. The data preparations techniques were used to aid the selection of the final variables for the 2011 OAC. Once the final variables were selected, different combinations of the same techniques were used to create multiple datasets. This facilitated the identification of the dataset that created the optimum clustering solution.

Section 6.6 details the processes used to aid selection of the final list of variables for the 2011 OAC. A number of techniques were used to help guide which of the initial list of variables should be retained. Section 6.7 explains how the optimum rate calculation, transformation and standardisation techniques for the final variables were selected to create the 2011 OAC. Section 6.8 gives an overview of the potential clustering processes

that could have been used with the 2011 OAC dataset. It also explains how the selected clustering method was used to create the structure of the 2011 OAC and helped provide the names and descriptions for the groups created. Additionally this section details the processes involved in ensuring that the clusters created were optimal, why this was important and the steps taken to make the procedure as reproducible as possible. Finally, Section 6.9 summarises the path taken in creating the methodology for the 2011 OAC and how it can be considered an evolution of the methodological approach taken with the 2001 OAC.

6.2. Cluster Analysis

Vickers (2006) states that "area classifications are created by the clustering of geographical entities with the use of cluster analysis" (p. 43), the basis of which is the ordering of a large and complex multidimensional dataset. The 2011 Area Classification for Output Areas (2011 OAC) can therefore be defined as seeking to bring similar objects, in this case Output Areas (OAs) or Small Areas (SAs), together to form distinctive groupings. Cluster analysis, unlike other statistical procedures, is a technique derived for data exploration (Vickers, 2006), which Everitt et al. (2011) describe as a "a convenient method for organising a large data set so that it can be understood more easily and information retrieved more efficiently" (p. 3).

Clustering, for the most part, is an unsupervised process (Kovács et al., 2005; Hasan et al., 2009). As the make-up of the clusters is not predefined, the evaluation of a final cluster solution can be difficult. While the aim of clustering is to group similar objects together, the degree to which this can be achieved will vary from cluster to cluster. Berry and Linoff (1996) state that an optimum cluster process should result in; 1) compact clusters, with the objects in each group being as similar in characteristic as possible, and 2) the highest possible separation in characteristic between the different clusters. The reality of clustering is that the variance in characteristics between objects within a cluster can be large, and these variances can be larger than those between the individual clusters. This can create an element of uncertainty within clustered groups, as objects within a single cluster may not be as similar as they first appear. As mentioned in Section 5.2, Slingsby et al. (2011) have visualised this phenomenon using the 2001 OAC. They demonstrated how OAs assigned to a cluster vary in how close they are to its centroid and therefore vary in how much they conform to the average characteristics of that cluster. This means that some OAs have characteristics in common with more than one

cluster, albeit rather loosely. OAs could therefore easily fit into more than one cluster group, which may not be apparent in the application of standard cluster analysis outputs, which tend to suggest uniformity across clusters. The nature of clustering is that it can be a fuzzy process, where objects inherently comprise characteristics of multiple groups. This is especially evident when clustering is applied to area classifications (Slingsby et al., 2011). Therefore the 2011 OAC, like the 2001 OAC, has unavoidable fuzzy characteristics.

It is important to distinguish clustering and cluster analysis. Clustering refers to the particular cluster method applied to group the data. There are many of these methods available to choose from, and some are described in detail later in this chapter. Cluster analysis, while based on a clustering procedure, is the entire operation required to cluster a particular dataset. In the case of the 2011 OAC this begins with the selection of variables – a particular characteristic of a person, household or dwelling expressed as a numerical measure or a category (ONS, 2013e) – and ends with the analysis and critique of the final clusters. Milligan and Cooper (1987) explain that the process of cluster analysis follows a series of steps. Each of these steps require numerous, often subjective, decisions to be made. The interaction of these decisions has an impact on the final clusters produced and thus makes creating a geodemographic classification as much an art form as a science (Vickers, 2006). The impact of the suitability of the decisions made greatly depends on the intended purpose of the classification (Lorr, 1983).

Milligan (1996) acknowledges that in the creation of a new geodemographic classification, the accumulated experience from the creation of past classifications should not be ignored. However, it is vital that a certain degree of freedom from previous work exists to guarantee a level of autonomy. Milligan (1996) originally stated the seven steps that form cluster analysis. These steps, derived from a number of studies, were further expanded to nine stages by Everitt et al. (2011). These are outlined below, providing a summary of the steps taken to create the 2011 OAC.

Stage 1: Clustered objects.

i) **The objects should be representative of the cluster structure believed to be present.** For the 2011 OAC, OAs and SAs were used as objects.

ii) **The objects should, if possible, give full geographical coverage.** The OAs and SAs in the 2011 OAC provide full and continuous coverage of the UK.

iii) **If required, the objects should be sampled to form an accurate representation of the population as a whole if generalisation is required.** This would only have been relevant for the 2011 OAC if sample based surveys had been included.

Stage 2: Selected variables.

i) **A variable should be reflective of the measurements taken.** The 2011 OAC achieves this by only using variables that are available at the smallest Census geography level.

ii) **A variable should only be included if there is good reason to believe it will add definition to the clusters.** For the 2011 OAC, it was vital to include variables that demonstrated spatial variation in their distribution. Without this spatial variation it would not have been possible to form unique clusters and areas would have been indistinguishable from each other.

iii) **Variables that do not help differentiate clusters should be excluded if possible.** The method used to select variables for the 2011 OAC was designed so that variables which demonstrated a uniform distribution across the UK were less likely to be included (see Section 6.6.5). Additionally, highly correlated variables were removed where possible to reduce redundancy in the dataset (see Section 6.6.1). This method removed the risk of masking unusual patterns within the dataset.

Stage 3: Missing values.

i) **When the proportion of missing values is low, imputation of the raw dataset may be acceptable.** This was not a consideration for the 2011 OAC as the data source utilised the 2011 UK Census which is the most complete enumeration of the UK's population (ONS, 2013a).

ii) **Alternatively, the elements in a similarity or dissimilarity matrix can be imputed using only variables that are present.** This was not necessary for the 2011 OAC due to the use of only Census data, as discussed in the previous point.

Stage 4: Variable standardisation.

i) **Standardisation of any dataset is not a requirement, instead it is the choice of those performing the cluster analysis, and if they believe it to be necessary.** Due to the variability of values between counts and densities, the 2011 OAC dataset was standardised (see Section 6.5.3).

ii) **Range standardisation produced good clusters (Milligan and Cooper, 1988) and should be considered an alternative to standardisation methods using standard deviations, such as z-scores.** This reflects the importance of testing different standardisation methods. Although the 2001 OAC utilised range standardisation (Vickers and Rees, 2007), it was one of several options considered.

iii) **The standardisation of variables is not necessarily always indicated in documentation that accompanies the final clustering result and can sometimes be misleading.** The decisions made during the construction of the 2011 OAC have been fully documented to reduce any ambiguity in the methodological process.

Stage 5: Proximity or distance measure.

i) **There are few guidelines; however knowledge of the context and type of data may suggest a suitable measure to use.** An understanding of the data used in the creation of the 2011 OAC formed an important part in the selection of the most appropriate proximity measure (see Section 6.8.2).

ii) **Proximity measures can be recorded in terms of similarity or dissimilarity.** An example of a similarity measure is a Pearson correlation coefficient, where the larger the value, the more similar two objects are likely to be. An example of a dissimilarity measure is the squared Euclidean distance dissimilarity measure; in this case the larger the value, the more dissimilar two objects are likely to be. The most suitable measure for the 2011 OAC was dependent on the clustering method used.

Stage 6: Clustering method.

i) **The method must recover the types of clusters suspected to be present.** This criterion was difficult to fulfil for the 2011 OAC, as any large population can be divided into numerous types of groups. A clustering method which divided the population into representative clusters was therefore used (see Section 6.8.3).

ii) **The method used should be robust enough to cope with a large dataset and be insensitive to outliers.** This was important for the 2011 OAC as it was clustered from a database containing multiple variables.

iii) **The method must be available in a software package.** The free command line program R (R Development Core Team, 2011) was used for the majority of the cluster analysis for the 2011 OAC. This program allows the script used to cluster to be made freely available and could therefore be easily reproduced.

Stage 7: Number of clusters.

i) **One of the most difficult decisions in cluster analysis; it is also the decision that defines the structure of a geodemographic classification**. The cluster numbers of the 2001 OAC (7 Supergroups, 24 Groups and 51 Subgroups) were not explicitly recreated; instead they were used as a guideline for the 2011 OAC (see Section 6.8.4).

ii) **There are several techniques that can support the selection of the number of clusters, although they can often be contradictory.** These techniques were found not to be suitable for use on the 2011 OAC due to the size of the database used for clustering.

iii) **If a decision cannot be made between two solutions, then the solution that results in the larger number of cluster should be used.** Any decision made regarding the number of groups in the 2011 OAC was made to favour the option which resulted in the most clusters.

iv) **It should be acknowledged that there may not actually be any clusters present within the data.** The creation of the 2001 OAC and the many commercial geodemographic classifications that exist for the UK would suggest that clusters can always be identified in large datasets concerning general population characteristics.

v) **There is not always an optimum number of clusters in a dataset. The most important factor when deciding upon cluster numbers is how useful these groupings are in the context of the classification.** The final number of clusters in the 2011 OAC was required to fulfil the criteria of representing the general characteristics of different population groups.

Stage 8: Replication and testing.

i) **Running the cluster analysis multiple times to guarantee a stable solution.** As R was used to perform the majority of the cluster analysis, the same clustering algorithm was run several thousand times to ensure a stable solution for the 2011 OAC (see Section 6.8.6).

ii) **Perturbation of the clustered dataset by omitting or slightly changing particular variables.** This allowed the identification of 2011 OAC variables which had the greatest and least impact on the overall cluster solution (using the method outlined in Section 6.6.3).

Stage 9: Interpretation.

i) **Interpretation of the results in the context of the applied problem and an assessment of whether the solution adequately meets the needs of the investigation.** The extent to which the 2011 OAC divided the population into robust and distinct groups was assessed.

ii) **This may require graphical representation, such as maps, and descriptive statistics. Proximity measures could also be used to better understand areas that have characteristics of more than one cluster.** These outputs were requested by the respondents to the 2011 OAC user engagement (See Section 4.3.2) and as such formed integral outputs of the 2011 OAC.

iii) **Standard statistical tests may be inappropriate to compare the variations between clustered variables across different clusters.** Validation of the 2011 OAC involved more than standard statistical tests and is discussed in Chapter 8.

These stages offered a thorough outline of the process of effective cluster analysis and provided a clear template for constructing the 2011 OAC.

6.2.1. Cluster Analysis and the 2011 OAC

Cluster analysis is at the heart of any geodemographic classification. Following the 2011 OAC user engagement (see Section 4.3.2), the creation of the 2011 OAC was guided by the 2001 OAC. The 2011 OAC cannot however be considered a carbon copy of what has gone before. Developments in computer processing since the creation of the 2001 OAC have been utilised to create a new and fully reproducible methodology which is fully documented.

Open geodemographic classifications strive to be as transparent as possible. The subjective and non-scientific aspects of geodemographics make this challenging, as not all will agree with decisions taken. However, provision of the resources to better understand the processes involved should be a priority in the creation of any open classification. The importance of documenting decisions was recognised by Vickers (2006) by citing Milligan (1996). An open classification provides an opportunity for users to actively engage with the entire process and further expand the classification if desired. The facility for others to critically evaluate, analyse or find other ways to actively engage with the process of creating a classification is an important tool in enhancing the understanding of that classification, and deciding whether it is appropriate for a particular task. Vickers and Rees (2007) documented the entire process of the creation of the 2001 OAC. This resource is available to anyone who wishes to use it and allows users to reproduce the 2001 OAC.

The approach taken by open classifications is not adopted by commercial organisations which create geodemographic classifications to sell to a wide range of customers (CACI, 2013c; Experian, 2013). They promulgate a 'black box' (Longley and Singleton, 2009) approach, meaning that the specific details of their data inputs and methodologies remain unknown (Harris et al., 2005). A detailed critique and wider understanding of how these classifications are formed can therefore not take place, which can be a desirable quality for users who do not require an understanding of the underlying process. Whilst this approach has obvious commercial advantages, it means that any open classification can never be truly thought of as equal to the commercial alternatives as comparison between the two is compromised (Brunsdon et al., 2011). To cater for the wide range of potential users, the 2011 OAC therefore needs to provide for both those who are only interested in the final classification *and* to those who would wish to understand or reproduce the whole classification. Open geodemographics, and especially the 2011 OAC, is not seeking to directly compete with any commercial product.

Instead, focus should be placed on highlighting the unique features that can only exist in an open and transparent environment.

6.3. Overview of the 2011 OAC Methodology

Creating the 2011 OAC involved the utilisation of the cluster analysis steps outlined by Milligan (1996) and Everitt et al. (2011). The steps provide a good base for any cluster analysis, but lack detail with regard to specific applications. Every geodemographic classification is unique, and consequently, each classification will have variations that do not implicitly follow the guidelines published in academic literature.

A total of 232,296 objects were clustered: 181,408 2011 OAs in England and Wales, 46,351 2011 OAs in Scotland and 4,537 SAs in Northern Ireland. A diagrammatic overview of the process involved in creating the 2011 OAC is shown in Figure 6.1. The majority of the operations undertaken were performed in the command line program R (R Development Core Team, 2011). This allows the scripts to be used by other users, and with the addition of full documentation allows the 2011 OAC to be a fully reproducible classification. A detailed explanation of the steps can be found in Sections 6.4 to 6.8.

6.4. Selecting Variables

A key output of the 2011 OAC user engagement was for the new small area classification to be based solely on 2011 UK Census data (see Section 4.4.1). The 2011 UK Census represents the most complete and reliable socio-economic dataset in the UK, with data available at the smallest OA and SA Census geography level. This allows for a geodemographic classification to be constructed at the smallest scale possible using only a free and easily accessible data source. It was for similar reasons that Vickers and Rees (2007) used only 2001 UK Census data to create the 2001 OAC.

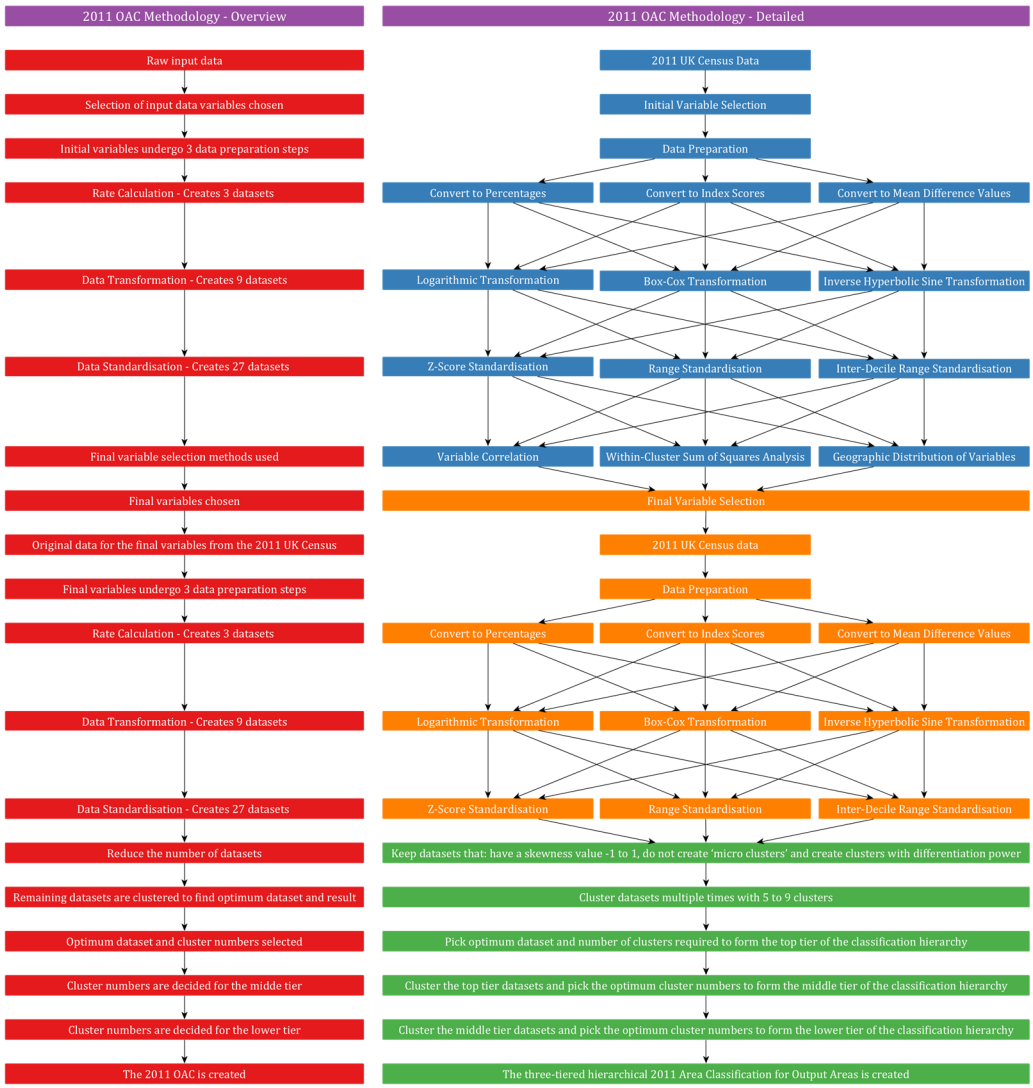


Figure 6.1: Overview of the 2011 OAC methodology

There was potential for additional Open Data sources to be utilised in conjunction with the Census data, indeed this was recognised by some of the respondents of the 2011 OAC user engagement (see Section 4.3.2). However, this was subsequently discounted for a number of reasons. Firstly it was felt that the current lack of Open Data available at the OA or SA level would create too many compatibility issues with the Census data. The rationale behind this is that the current Open Data sources with complete UK coverage are disseminated at a higher spatial level such as Lower Layer Super Output Areas (LSOAs), which would need to be aggregated down to the OA or SA level for use in the 2011 OAC.

Vickers and Rees (2007) similarly discounted other non-OA or SA level datasets due to the inherent uncertainty and reliability issues caused by aggregating datasets. The second issue is that the only Open Data sources currently available at OA or SA level do not have full UK coverage. As the objective of the 2011 OAC is to ultimately create a UK-wide classification, any data source which does not have full UK coverage has to be discounted to guarantee that the classification is created using the most robust, and spatially continuous, set of variables possible. It should be noted that whilst Open Data have been discounted for the UK-wide 2011 OAC, it should be considered for those using the methodology on a smaller geographical area and dataset.

6.4.1. Initial Variable Selection

Outputs of the 2011 UK Census at the OA and SA level were released in stages by the Office for National Statistics (ONS) for England and Wales, National Records of Scotland (NRS) for Scotland and the Northern Ireland Statistics and Research Agency (NISRA) for Northern Ireland. The outputs that were the most useful for creating the 2011 OAC were those provided in a univariate format. For the 2011 UK Census data, these have been branded by the statistical bodies as 'Key' and 'Quick' statistics. The ONS released 35 tables as Key Statistics and 73 tables as Quick Statistics for England and Wales, the NISRA released 45 tables as Key Statistics and 58 tables as Quick Statistics for Northern Ireland, and by December 2013 the NRS had released 34 tables as Key Statistics and 59 tables as Quick Statistics for Scotland. The combined dataset for England and Wales contained 2,139 variables, in Scotland there were 1,326 and in Northern Ireland 1,378. Only variables that were consistent across the whole UK were considered for use in the creation of the 2011 OAC. However, within this reduced dataset there were numerous cases of variable duplication, for example tables KS101EW and QS101EW contained

identical data regarding the usual resident population of England and Wales. Key Statistics are designed to highlight key attributes, albeit in an aggregated format. For example, age bands in Quick Statistics are expressed by single year of age but for Key Statistics they are aggregated to pre-defined bands that have been deemed as useful break points for end users. Quick Statistics can therefore be seen as providing more detailed information about the topics that each table provides. It was therefore appropriate to consider both Key and Quick statistics as a single dataset for the purposes of selecting the best possible variables for the 2011 OAC.

The large number of potential variables available to create the 2011 OAC provided an ample selection of different attributes of the population for consideration. The overall aim for the 2011 OAC's variables was to have the smallest selection which represents the main components of the 2011 UK Census (Bailey et al., 1999, 2000). To aid in this, like the 2001 OAC before, these potential variables were categorised into five domains: Demographic Structure, Household Composition, Housing, Socio-Economic and Employment (Vickers et al., 2005). Retention of the same domains as the 2001 OAC allowed for similar variables to be used in the 2011 OAC, without restricting the final variable selection to a mirror image of the previous classification. This accommodates a key output from the 2011 OAC user engagement (see Section 4.3.2), that a like-for-like replacement of the 2001 OAC was not desirable, but at the same time should bear some similarities.

The inclusion of every variable for all the OAs and SAs in the UK would have created an input database with very high dimensionality, raising significant issues in the assembly and interpretation of outputs. One method by which this issue has traditionally been tackled is through Principal Components Analysis (PCA), where individual variables are reduced to a series of linearly uncorrelated component scores. Historically this has been necessary because the clustering of large multidimensional datasets was either unachievable or would take too long to process. Current computational power means that clustering large multidimensional datasets can now be completed in hours or even minutes. Vickers (2006) discounted the utilisation of PCA for the 2001 OAC because clustering from an entire dataset was seen as favourable, and advances in computational power meant that actual data points, rather than principal components, could be used. The PCA method can however be useful in an exploratory sense, for example showing the discriminatory power of selected variables and being potentially useful for removing variable redundancy. Others have however criticised PCA for its tendency to remove any

non-linear relationships from the data (Harris et al., 2005). The PCA method was not used for the 2011 OAC and the importance of individual variables were given the greatest consideration.

Given the above, a process by which important dimensions can be selected was required. As has historically been the case with previous Census based classifications, variables were therefore initially selected manually based on the premise that only those considered useful in creating a general purpose geodemographic classification were warranted – i.e. are reflective of those attributes that may be important in those potential domains of use; and additionally, those that a priori are known to be important in driving patterns of socio-spatial structure (Webber and Craig, 1978). The findings of the 2011 OAC user engagement, discussed in Section 4.4, also helped to guide this process, especially in relation to what users considered to be key uses of the classification. As the utilisation of both Key and Quick statistics lead to duplication, it was necessary to decide which version of the variables to use. It was important to remember the principles stated by Bailey et al. (1999, 2000) to keep the total number of final variables to a minimum, whilst still representing the main dimensions of the Census data.

This list of variables selected after this initial variable selection, known as 2011 OAC prospective variables, was reduced further into the final list of variables after they had been prepared using the procedures outlined in Section 6.5 and gone through the processes outlined in Section 6.6. The processes used in selecting the final variables are identified in the blue boxes in Figure 6.1.

6.5. Data Preparation

This section outlines the steps taken to prepare the 2011 UK Census data for clustering. The methods outlined were used firstly in the reduction of the prospective variables into a final variable dataset, and secondly in the preparation of the data that were clustered to create the 2011 OAC.

As raw data counts for different variables require different denominators, they are not comparable in their original format (Walford, 2013). The process of rate calculation, transformation and standardisation on the prospective variables dataset was therefore carried out first on the raw data. Correlation and within-cluster sum of squares (WCSS) analysis was then utilised to aid in the selection of the final list of variables. The explicit

criterion of solely using data that had undergone these processes ensured a level of consistency in the construction of the 2011 OAC.

For both variable selection and final clustering, the 2011 OAC tested a greater number of rate calculation, transformation and standardisation techniques than its 2001 OAC predecessor. This allowed for a more optimal selection of procedures to be identified and utilised for clustering of the final classification. The procedures that are outlined below test a number of rate calculation, normalisation and standardisation procedures; which are then evaluated in relation to their impact on final classification assignments. As such, only one combination of procedures was used in the final clustering, but in order to identify the optimum process all combinations were evaluated.

6.5.1. Rate Calculation

The 2011 OAC considered three rate calculation techniques. The first method entailed conversion of the raw data into percentages. The majority of the variables in the 2011 UK Census are provided as counts. These counts can however have different base categories. For example, counts that relate to the whole population, such as age, require the total population for each OA or SA as the denominator. Whilst counts that relate to the working population only require the total population of 16 to 74 years olds for each OA or SA. The denominator is therefore not constant between variables. There are however some variables within the Census data, such as those relating to area or density, that cannot be transformed into percentages. Such variables were therefore left untouched at this stage.

Calculating percentages are defined as:

$$\frac{\Sigma \text{ target OA or SA}}{\Sigma \text{ OA or SA denominator}} * 100 \qquad (6.1)$$

The second rate calculation procedure tested was index scores. Index scores show how overrepresented the characteristics of the target group are relative to the base UK population – in the case of the 2011 OAC this is the percentages of each variables population in relation to its denominator. For every variable supplied as a count index scores were calculated as follows:

$$\frac{\%\text{ population in target OA or SA}}{\%\text{ population in base denominator OA or SA}} * 100 \quad (6.2)$$

The final rate calculation procedure tested with the 2011 OAC was the mean difference. This is the difference between the actual values for all OA and SA variables and the expected values and can be defined as:

$$\left(\frac{\Sigma\ \text{OA or SA}}{\Sigma\ \text{OA or SA Denominator}} * \Sigma\ \text{OA or SA Total Population}\right) - \Sigma\ \text{OA or SA} \quad (6.3)$$

This method can be used to identify variables which exhibit the most deviation away from their respective 'average' characteristics. This allows the extent to which OAs and SAs conform to a particular variable's national average to be identified, and potentially aid delineation of the population.

These initial rate calculation procedures generated three distinct datasets from the raw 2011 UK Census data.

6.5.2. Data Transformation

The three datasets detailed in Section 6.5.1 were analysed to explore the extent to which attribute data were normally distributed. It can be argued that highly skewed data could lead to poor assignments if clustered with algorithms that are optimised to find spherical groupings of cases with similar attributes (e.g. k means). There are however differing views on the optimal method, Harris et al. (2005) describe how in commercial geodemographics skewed (or 'misbehaving') attributes can be controlled through down weighting of the variables. However, the decision was taken not to weight attributes of the 2011 OAC, given the inherent subjectivity to these choices, and a mismatch with the 2001 OAC methodology (see Section 6.6.6). A modification of this approach was adopted by Singleton and Spielman (Forthcoming), where data were not normalised, but also, no weightings were applied. They argue that weightings are subjective, but through global normalisation there is potential that interesting local patterns and interactions are smoothed away. To an extent, this mirrors some of the discussions of Harris et al. (2005) around the use of data reduction techniques such as principal component analysis. These views contrast with those of Vickers et al. (2005) who discuss that a large number of variables with skewed distributions can have a negative impact on final cluster solutions by giving too much prominence to outliers in a dataset. Based on their experiences of

building the 2001 OAC, a large number of outliers existed towards the high end of the value scale with population density being a particular issue. As such, utilisation of a transformation technique can reduce any skew in the distribution of the dataset caused by these high value outliers. Furthermore, as identified in the user engagement (see Section 4.3), there are constraints in that the 2011 OAC should have a passing resemblance the 2001 OAC methodology, and as such, the remainder of this section explores the issue of normalisation by exploring the log method implemented in 2001 OAC and a series of potential alternatives.

Three different forms of transformation were tested on each of the datasets created in the previous section:

1. Log
2. Box-Cox
3. Inverse Hyperbolic Sine

Log and Box-Cox both require values to be positive and greater than 1 to transform. There are different ways of managing this issue, however, most common is for a constant to be added to all values before implementing the normalisation method. For the percentages and index score datasets, zero is the lowest value possible, and as such, by adding 1 to the values, this allowed the techniques to function (Osbourne, 2002). The use of the mean difference method was however constrained by the presence of negative values being produced for each variable. As such, these values were re-ranged between 1 and 2 as if left unmodified the normalisation procedures could not have been performed.

Log transformations are defined by Aitchison and Brown (1957) as "the distribution of a variate whose logarithm obeys the normal law of probability" (p. 1). The purpose of the transformation is to make the data conform to the lognormal law of error for inferential purposes. This is achieved by artificially reducing the amount of variance to that of the normal distribution (Leydesdorff and Bensman, 2006). In essence it makes the differences between the larger values less significant, while simultaneously making the differences between the smaller values more significant. A disadvantage of this approach is that the transformation is fixed across a dataset without sensitivity to different distributions that may appear between variables. An alternative method more sensitive to these issues is the Box-Cox transformation which can be defined as:

$$x_i(\lambda) = \begin{cases} x_i^{\lambda} - 1)/\lambda & (\lambda \neq 0) \\ \log(y) & (\lambda = 0) \end{cases} \quad (6.4)$$

An exponent, lambda (λ), transforms a variable (x) into a Normal distribution (Box and Cox, 1964). Multiple values for lambda are tested, and the one that produces the most normal result is selected. There are numerous tests of normality that can be applied, however, for the implementation here the common Shapiro-Wilk test was used. If in the selection of lambda, a value of 0 is identified, mathematically, this would return the same result as log. As such, the implementation of the Box-Cox technique calculates a separate lambda value for each variable. The extent to which a variable is transformed will depend on how skewed its distribution is, rather than a global skewness value (Dag et al., 2014).

As noted in discussion of the previous two techniques, where zero values are present, a constant is required to enable transformation. However, an alternate approach without such constraints is available in the inverse hyperbolic sine transformation which is defined as:

$$\log(\, x_i + ({x_i}^2 + 1)^{1/2}) \quad (6.5)$$

where

$$x = \text{a variable} \quad (6.6)$$

The inverse hyperbolic sine (IHS), proposed by Johnson (1949) shares similarities with the standard log transformation, except that it is not defined at zero (Burbidge et al., 1988). As a result, the technique is favoured when transforming wealth datasets, where a large number of values are either zero or negative (Pence, 2006).

As discussed earlier, the normality of attribute data can have an impact on the outcome of clustering, and as such, the transformation technique chosen can impact any output representations. As such, it was important to identify the impacts of the different transformation techniques tested upon the final assignment of areas into categories post clustering. This builds on the work of the 2001 OAC, which only utilised a standard log transformation. Such evaluation is especially important in a general purpose classification like the 2011 OAC as this needs to be driven by the choice of variables,

rather than the effects that different transformation techniques may have on the resulting classification.

6.5.3. Data Standardisation

To ensure that all variables are measured on the same scale in the cluster analysis, and as such have the same influence, a process of data standardisation is required. As with the 2001 OAC, three data standardisation techniques were explored: z-scores, range, and inter-decile range. Each of these three methods were applied to the nine converted and transformed datasets described in Sections 6.5.1 and 6.5.2. The final choice of method (normalisation, plus standardisation) was evaluated in the context of producing the best geodemographic representation of the study population.

6.5.3.1. Z-score standardisation

Z-scores are one of the most widely used approaches of variable standardisation (Adnan et al., 2010). Z-scores standardise the original distribution of the data so the mean (x_{mean}) becomes 0 and the standard deviation (σ_x) becomes 1, quantifying the original values (x_i) in terms of the number of standard deviations they are away from the mean.

A z-score can therefore be defined as:

$$Z_i = \frac{x_i - x_{mean}}{\sigma_x} \tag{6.7}$$

where

$$\sigma_x = \sqrt{\frac{\sum_i (x_i - x_{mean})^2}{n}} \tag{6.8}$$

and

$$\text{n = The number of values (e.g. the total number of OAs and SAs)} \tag{6.9}$$

This method can emphasise the effect of outlying observations in the datasets, which may serve to highlight interesting patterns within the data: however, such observations can also adversely influence some clustering algorithms.

6.5.3.2. Range standardisation

Range standardisation compresses the values in a dataset into the range of 0 to 1. It compares each value of a variable (x_i), to the minimum value (x_{min}) of that variable. This is divided by the difference between the minimum value (x_{min}) and the maximum value (x_{max}) of the variable. The lowest value(s) in the dataset will be assigned 0 and the highest value(s) assigned 1.

Range standardisation can therefore be defined as:

$$R_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{6.10}$$

A result of this method is that any outliers present within the dataset get compressed to fit within this 0 to 1 interval. As such, interesting patterns within the dataset can be omitted, a potential negative aspect of using this method. The 2011 UK Census data contain numerous zero counts; these are therefore assigned the lowest value, 0. Any outliers in the dataset are therefore found at the other end of the scale. To accommodate these outliers, the majority of the values are actually found within a much smaller range. The range standardisation method was used for the ONS 1991 classification of Local Authorities (see Wallace and Denham, 1996) and for the 2001 OAC (see Vickers and Rees, 2007).

6.5.3.3. Inter-decile range standardisation

The inter-decile range standardisation method is a variation of range standardisation. The data are standardised over a smaller range, between the 90th percentile and the 10th percentile. This aims to reduce the impact of outliers on the standardised data. The median (x_{med}) for each variable is subtracted from the variable value (x_i). This is then divided by the distance between the 90th percentile (x_{90th}) and the 10th percentile (x_{10th}).

Inter-decile range standardisation can be defined as:

$$D_i = \frac{x_i - x_{med}}{x_{90^{th}} - x_{10^{th}}} \tag{6.11}$$

Vickers and Rees (2007) found that this method gave too much weight to skewed variables and it was consequently not considered for use with the 2001 OAC. However,

it was still appropriate to consider the method as a standardisation option for the 2011 OAC due to differences in the variable selection compared to the 2001 OAC, and differences between the two Census datasets.

6.6. Final Variable Selection

The process of selecting the final variables took into consideration a combination of two key requirements. Firstly, that the 2011 OAC was to be a general purpose geodemographic classification. This meant including a collection of variables that reflected the general characteristics of as much of the UK's population as possible, but also had geographic variance in order for distinct clusters to form. Secondly, that only the minimum number of variables required were selected in order to reduce the impact of co-linearity. The process of reducing the prospective variables to a list of final variables to cluster required numerous steps with the overarching aim to maintain only those variables that were likely to be the most important to the 2011 OAC. Table 6.1 outlines the 27 unique datasets created from the different combinations of rate calculation, transformation and standardisation techniques described in Section 6.5. The processes described in this section were used on each of these datasets, with the outputs of these techniques aiding the final variable selection.

Vickers (2006) highlighted the problem of vague or uncertain variables when selecting variables for the 2001 OAC. For example, the assessment of whether housing with no residents was either 'vacant' or a 'second residence/holiday accommodation' was made by the enumerators who delivered the 2001 UK Census forms to households. Vickers (2006) suggested that the enumeration method led to less housing being categorised as a second residences or holiday accommodation. In England the 2001 UK Census recorded 811,000 vacant properties, 7% less when compared to Council Tax returns (Bolton Metropolitan Borough Council, 2005). Consequently, variables derived from these data have an increased level of uncertainty with regards to accuracy. For the 2011 UK Census this problem was partially addressed by a change in the distribution method of the Census forms. Forms were delivered by post, or could be completed online. The respondent, therefore, completed all answers from the Census, with no subjective decisions being made by enumerators. This creates more certainty that the responses to the 2011 UK Census are a true reflection of the population, thereby eliminating the need to be cautious about using certain variables.

Table 6.1: Datasets created from applying the rate calculation, transformation and standardisation procedures to the 2011 OAC variables

Step 1	Step 2	Step 3
Percentages	Percentages, Box-Cox	Percentages, Box-Cox, Z-Scores
		Percentages, Box-Cox, Range
		Percentages, Box-Cox, Inter-Decile Range
	Percentages, Log	Percentages, Log, Z-Scores
		Percentages, Log, Range
		Percentages, Log, Inter-Decile Range
	Percentages, Inverse Hyperbolic Sine	Percentages, Inverse Hyperbolic Sine, Z-Scores
		Percentages, Inverse Hyperbolic Sine, Range
		Percentages, Inverse Hyperbolic Sine, Inter-Decile Range
Index Scores	Index Scores, Box-Cox	Index Scores, Box-Cox, Z-Scores
		Index Scores, Box-Cox, Range
		Index Scores, Box-Cox, Inter-Decile Range
	Index Scores, Log	Index Scores, Log, Z-Scores
		Index Scores, Log, Range
		Index Scores, Log, Inter-Decile Range
	Index Scores, Inverse Hyperbolic Sine	Index Scores, Inverse Hyperbolic Sine, Z-Scores
		Index Scores, Inverse Hyperbolic Sine, Range
		Index Scores, Inverse Hyperbolic Sine, Inter-Decile Range
Mean Difference	Mean Difference, Box-Cox	Mean Difference, Box-Cox, Z-Scores
		Mean Difference, Box-Cox, Range
		Mean Difference, Box-Cox, Inter-Decile Range
	Mean Difference, Log	Mean Difference, Log, Z-Scores
		Mean Difference, Log, Range
		Mean Difference, Log, Inter-Decile Range
	Mean Difference, Inverse Hyperbolic Sine	Mean Difference, Inverse Hyperbolic Sine, Z-Scores
		Mean Difference, Inverse Hyperbolic Sine, Range
		Mean Difference, Inverse Hyperbolic Sine, Inter-Decile Range

6.6.1. Variable correlation

It can be expected within any large multidimensional dataset there is always likely to be an element of correlation between variables. As a general rule for geodemographic classifications, highly correlated variables are not desirable (Ojo et al., 2012). They act to weight classifications by giving variables that highly correlate an increased prominence. However, some highly correlated variables were retained for use in the 2001 OAC (Vickers et al., 2005). This was done to add predictive and descriptive power to the classification, enabling behaviours not included within the initial specification to be predicted because of the high correlation with the remaining variables. Vickers (2006) states that examining and assessing the correlations between variables is a more transparent way of reducing redundancy within a dataset compared to methods like PCA.

Figure 6.2 presents an example that illustrates one possible way of interpreting the correlation coefficient, but as Gunesh (2005) states, this should only act as a guideline and there is no particular value when the correlation switches from moderate to strong. As there is no stated rule for a high correlation coefficient, an arbitrary decision needs to be made on which variables will be deemed to be 'highly' correlated. Ideally, the threshold should be between +/-0.6 and +/-0.7, as this would allow the most highly correlated variables to be identified, and would leave the variables with weaker correlations untouched.

The larger the dataset, the greater the number of correlated variable permutations that are possible; with the 41 variables of the 2001 OAC having 820 unique permutations. Consequently, a variable can be highly correlated with multiple other variables. Highly correlated variables remained in the 2001 OAC as there was no practical way of removing them without removing variables that were important to the overall classification. There are three options available after identifying highly correlated variables: remove the variable from the final selection; group it with other variable(s) to create a new composite variable; or ignore it. Documentation to better understand the relationship between the 2011 OAC's final variables has been produced. Material such as correlation matrices and a statement of the threshold figure that was used to determine whether variables were considered highly correlated have been produced. This aids the understanding of why some variables were removed, why others remained and how the remaining variables interact.

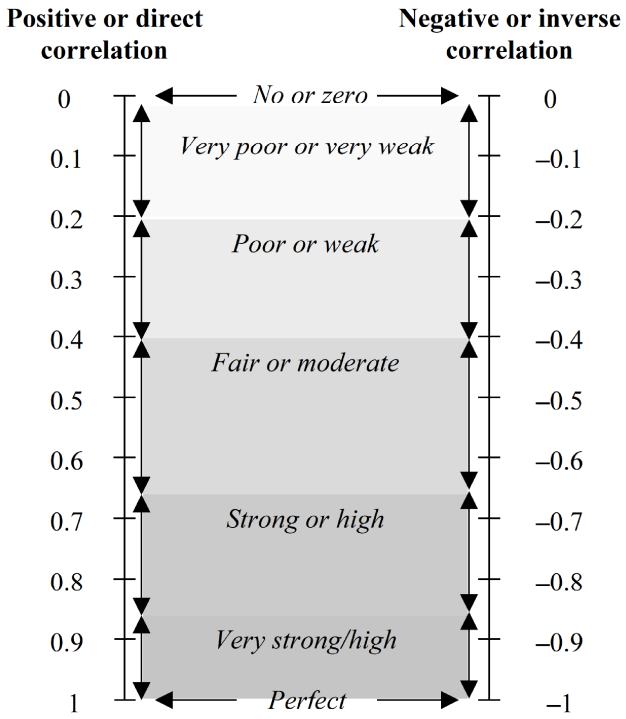


Figure 6.2: Interpretation of correlation coefficients

Source: Gunesh (2005)

6.6.2. Composite variables

The second stage of variable selection was the derivation of composite variables. These are formed from variables that are related and show similar patterns; for example, separated and divorced people were merged into a single variable for the 2001 OAC (Vickers et al., 2005). This meant that highly correlated variables that represented an important aspect of the population were joined. This removed the redundancy, and artificial weighting in the classification caused by having two variables with similar characteristics. Other composite variables were created from individual variables that only represented a small percentage of the total population. If left, these individual variables would have had minimal impact on the final clusters due to their high frequency of zero values. Grouping such variables together into a single variable meant it represented a greater share of the population, thereby having a greater impact on the final clusters. The creation of this type of composite variable is a more subjective process, and largely dependent on personal discretion. An example of this from the 2001 OAC is the combination of the 'Indian', 'Pakistani' and 'Bangladeshi' variables into a single composite variable (Vickers, 2006). Together, these variables represented a greater share of the population and therefore had a larger impact on the classification, although

they could have easily remained as individual variables. In some instances grouping such variables together was not appropriate. The rationale for this is while the influence of a variable across an entire population may be small; the areas where a concentrated amount is located can provided a key indicator of local social and demographic structure.

The creation of composite variables was therefore a more complex procedure than the identification of highly correlated variables which shared the same denominator, or variables that represented a small slice of society. Additional factors were considered and justified within the remit for the classification and the final selection should offer an overview of the general characteristics of as much of the population as possible.

6.6.3. Within-cluster sum of squares analysis

The ability to identify which variables had the greatest impact on a classification was an important consideration in variable selection. It is possible to analyse the relative impact each variable has by performing within-cluster sum of squares (WCSS) analysis. The WCSS value indicates how tightly clustered a particular dataset is. A smaller value signals that objects within each cluster are closer to their centroids, thereby increasing cluster homogeneity. In addition, a smaller WCSS value creates a higher between-cluster sum of squares (BCSS) value, meaning that the differences between individual clusters increases. For the purposes of variable selection, when discussing the WCSS value, this refers to the total mean value of all cluster WCSS values within a clustering solution. Removing any variable from an overall cluster solution impacts the WCSS value. If a variable is removed and there is a marked decrease in the WCSS value, this indicates that the final cluster solution would have more homogenous clusters without the inclusion of that variable, suggesting that the variable should be discarded. Only a slight decrease in WCSS value for a variable suggests that its impact on the homogeneity of the clusters is minimal and therefore should remain.

Figure 6.3 shows the results of the application of this technique to the 41 variables used to construct the 2001 OAC. Removal of variable 18, 'terraced housing', resulted in the greatest reduction of the total mean WCSS value. The classification thus appears to be sensitive to the inclusion of housing variables (16 to 21). These variables produce the five lowest WCSS values, indicating that there is a potential problem in summarising the different types of housing ownership found in the UK into three categories and rental property into two. These results indicate that although following the fundamental

principle of creating a classification with the smallest number of variables possible is useful (Vickers et al., 2005), it should be used with caution, and not lead to a compromise of the final cluster solution. Despite making the final clusters of the 2001 OAC less homogenous, the inclusion of the terraced housing variable was important to the overall effectiveness of the classification. It allowed for differentiation between different types of housing, which subsequently allowed clusters to delineate the population more effectively.

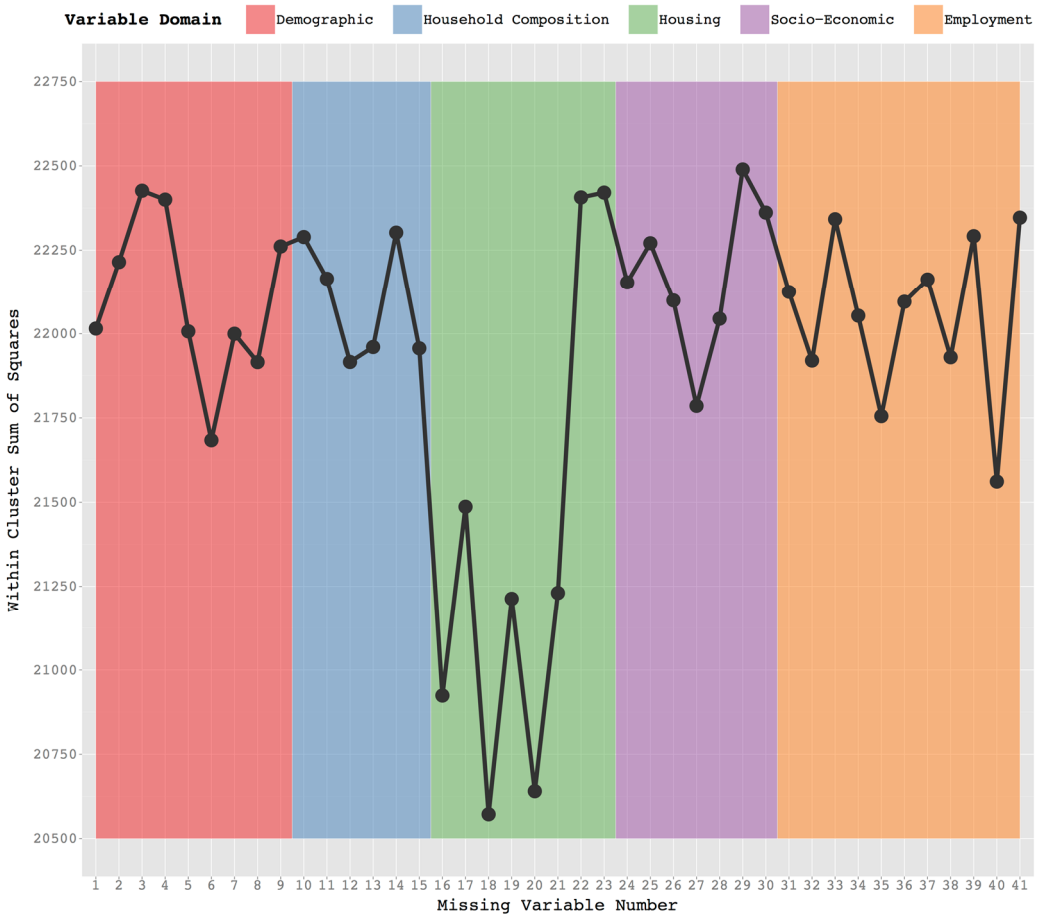


Figure 6.3: Missing variables WCSS values for the 2001 OAC
(See Table 2.2 for variable names)

The use of this type of WCSS analysis for the 2011 OAC was used to aid variable selection. The prospective variables were removed one at a time, whilst the remaining variables were clustered using the k-means algorithm (see Section 6.8.1.1) to produce 5 to 9 cluster solutions. A mean WCSS value for each missing variable was then produced and visualised (similar to Figure 6.3). The mean WCSS value provided an indication of how each variable impacted upon the homogeneity of the final clusters. Variables that had an adverse impact were more likely to be disregarded from the final variable selection. This however depended on each variable's perceived importance as a social or demographic indicator of the population, and as such were not automatically removed.

6.6.4. Skewed variables

Clustering algorithms function best when the data being used have a normal distribution (see Section 6.5.2). The skewed nature of a dataset can be evaluated at a global scale (i.e. looking at the dataset as a whole) or by each variable in turn. For the purposes of selecting the final variables it is advantageous to look at the skewed nature of each variable individually. Keeping individual variables with skewed distributions increases the possibility that a clustering solution will not be derived in the most appropriate way. The rate calculation, transformation and standardisation processes carried out prior to clustering attempted to ensure that each variable had an approximate Normal distribution (see Section 6.5). However, while data preparation reduced the impact of some of the skewed variables with the 2011 OAC, certain variables will always retain a skewed distribution.

There can be several reasons why a variable has a skewed distribution; for example they reflect a characteristic that is uncommon across the UK. Vickers (2006) highlighted the occurrence of communal establishments from the 2001 UK Census; 88% of OAs across the UK lacked any form of communal establishment, resulting in a skewed distribution for that particular variable. Skewed variables should not automatically be excluded however, as it can be a particular variable's presence or lack of, that is of significance to an area. In these circumstances a decision is made on the positive and negative impacts of including that variable. The skewed distribution of a variable was an important consideration for the 2011 OAC, but was considered to be less of a priority than selecting variables which best represented the social and demographic characteristics of the population as a whole.

6.6.5. Geographic distribution of population characteristics

Variables used to construct a geodemographic classification need to vary from one area to the next. The Supergroups of the 2001 OAC revealed only one cluster that represented the rural characteristics of the UK, while the other six were focused more on differentiating between urban areas. The formation of multiple clusters in urban areas is an indication of greater variation in the extent to which variables occur in these locations. Selecting certain variables that highlighted variation between areas, and in particular urban locations was therefore seen as a priority for the 2011 OAC.

Vickers (2006) examined the variation of ethnic groups across the UK for inclusion in the 2001 OAC. Citing research by Peach (1996), the distribution of White and Chinese populations were more homogeneous when compared to Black Caribbean, Black African, Black Other, Indian, Pakistani and Bangladeshi residents. As these populations had more spatial variation, they were included in the 2001 OAC, while the White and Chinese equivalents were not as they were perceived to be poor indicators of characteristics of a particular area. Ethnic variables however, can be some of the most geographically diverse variables. The release of the 2011 UK Census data has shown that London is the most ethnically diverse area, and Wales the least (ONS, 2012b). One of the most interesting results from the 2011 UK Census is that England and Wales as a whole is becoming more ethnically diverse, following previous trends for London. The White ethnic group accounted for 86% of the population in 2011, a decrease from 91.3% in 2001 and 94.1% in 1991 (ONS, 2012b). Therefore the inclusion of variables that appear integrated into the wider population were not automatically discounted. Their inclusion in a classification can aid in the distinguishing of clusters at lower spatial levels where concentrations of a variable that are not seen at a broader geographical scale become evident.

6.6.6. Variable weighting

Variable weighting is the process whereby certain variables are given increased prominence prior to clustering. Romesburg (2004) suggested that the weighting of variables is most appropriate when there is a specific goal for a classification. In such scenarios it would be possible to assess the impact of weighting, i.e. does it make a classification more effective for its particular goal? The weighting of variables for the 2001 OAC may have reduced the suitability of it for certain applications (Vickers et al., 2005). The classification was designed to encapsulate the general characteristics of as

much of the total UK population as possible at the small area level. The design of the classification is for utilisation in multiple applications, as such there was no guarantee that any weighting used would result in a beneficial solution to all.

The feedback received from the 2011 OAC user engagement exercise suggested that none of the current and former users found any negative impact from the lack of implicit weighting on how they used the 2001 OAC. Additionally, the feedback suggested that the majority of respondents wanted the 2011 OAC to remain as a general purpose classification.

Although Vickers et al. (2005) argue that weighting variables is just as likely to have a negative impact as it is positive for a general purpose classification, commerical alternatives, such as Acorn and Mosaic have adopted the technique (CACI, 2013a; Experian, 2013). There is however likely to be a metholdological rationale for its utilisation; without the weighting, certain variables could have too little discriminatory power in the final cluster solutions. Unlike the 2001 OAC and 2011 OAC, commercial classifcations use datasets with varying (and undocumented) coverage of the population and spatial area of the UK. Weighting is therefore likely required to ensure that the data used is refelective of the entire geographical region.

There remains a possibility that certain variables could have too much prominence in a general purpose classification using Census data. The use of weighting would counteract the inter-correlation that can still exist in the final selection of variables. The artificial weighting caused by this inter-correlation can have a potential detrimental impact on clustering solutions. However, it was concluded that including weighted variables in the 2011 OAC would have added too many unknowns into the classification due to the significant number of interactions within such a large dataset. The disadvantages of including weighted variables outweighed any advantages, and like the 2001 OAC, focus was instead placed on obtaining the best variable selection (Vickers, 2006).

6.7. Optimum data preparation techniques

As described in Section 6.5 and shown in Table 6.1, a result of testing multiple data preparation techniques for the 2011 OAC was the creation of 27 unique datasets produced from unique combinations of rate calculation, transformation and standardisation. The techniques described in Section 6.6 were used on the 27 different

datasets containing the initial variable selection to aid the final choice. All of the 27 datasets were given equal prominence to avoid the influence of any possible previous assumptions on the final classification.

The steps described in Sections 6.5 and 6.6 were then repeated once the final set of variables had been chosen. This produced another 27 datasets, all of which could have been used for cluster analysis and the creation of the 2011 OAC. Although the creation of the datasets themselves was not a time consuming process, performing cluster analysis and assessing the suitability of each to form the final classification would have been. As such, methods were used to reduce the 27 datasets containing the final variables to a smaller number that could subsequently be individually assessed.

As noted in Section 6.6.4 the skewed nature of the data has a significant impact on the final clustering solution. Datasets were only considered for clustering if they had an overall skewness value of between -1 and 1. This global skewness value, derived from combining the skewness values of each variable, allowed for the identification of datasets that were likely to contain a large number of variables that were not desirable for clustering.

This method did however have its limitations, as a skewness value does not indicate the modality of the distribution. This is less problematic for datasets with a unimodal distribution as it makes a skewness value easier to interpret. Visualising the skewness values allowed for the type of distribution to be identified. The presence of other types of distributions, such as bimodal, meant that alternative methods were necessary to help reduce the number of datasets to cluster. As the -1 to 1 range was an arbitrary threshold, this value could have been adjusted depending on how many datasets were left or discarded. This method was evaluated as being the most effective at giving a quantitative rationale for clustering certain datasets and discarding others.

In addition to solely discarding datasets which had skewness values outside of these thresholds, alternative methods were used to identify other datasets that could be removed from consideration. The clustering algorithm detailed in Section 6.8.3 was used to cluster the remaining datasets. Initially the only outputs evaluated were the cluster assignments, and whether a dataset had a tendency to create clusters with low numbers of OAs or SAs assigned to them (termed ‘micro clusters’). Datasets which resulted in this phenomenon were therefore discarded. To reduce the number of datasets further, those

which formed clusters with poor differentiation, where there was little to distinguish one cluster from another, were also removed. Finally, the cluster solutions of the remaining datasets were mapped. They were examined for a solution that looked the 'most right', guided by Benoit Mandelbrot's view that "the basic proof of a stochastic model of nature is in the seeing: numerical comparisons must come second" (Mandelbrot, 1982a, p. 581). Batty and Longley (1986) argued that this statement can also be applied to artificially generated phenomena. As such, the dataset which produced the result that looked the best while also utilising robust statistical components, such as homogenous clusters, was regarded as optimum. As a result the data preparation techniques used to create this dataset were also considered optimum for creating the 2011 OAC.

6.8. Clustering

Selecting a clustering method can be considered more of a subjective process than one based on statistical evidence. It is therefore essential to provide the rationale for the opted method in the creation of a geodemographic classification. This section describes the clustering techniques considered for the 2011 OAC.

6.8.1. Common geodemographic clustering techniques

There are numerous methods (or algorithms) that can be used to cluster a dataset. These methods can be grouped into four types (Jain et al., 1999; Kovács et al., 2005):

i) **Partitional Clustering:** Directly decomposes a dataset into a set of disjoint clusters. Typically the global criteria involve minimising some measure of dissimilarity within each cluster, while maximising the dissimilarity of different clusters (Kaski, 1997). This optimisation process is an iterative procedure.
ii) **Hierarchical Clustering:** Creates clusters recursively. They create clusters by either merging smaller clusters together, or splitting larger clusters into smaller ones.
iii) **Density-based Clustering:** Creates clusters based on density functions. A potential advantage of these algorithms is that they create arbitrarily shaped clusters.
iv) **Grid-based Clustering:** Summarises a dataset into a grid representation and subsequently merges grid cells to create clusters (Akodjènou-Jeannin et al., 2007).

An important aspect of the most prominent clustering methods (e.g., Ward's method and k-means) is that they produce virtual cluster centres (centroids). The centre of each cluster is a centroid that is defined by the variable means for that cluster. The centre of a cluster is therefore virtual in the sense that it typically does not correspond to any specific object in that cluster.

6.8.1.1. K-means

The k-means algorithm (MacQueen, 1967) is a form of partitional clustering that involves an iterative algorithm that operates on a fixed number of clusters (Van Laerhoven, 2001), or seeds (k), visualised in Figure 6.4. The seeds are randomly placed in multidimensional space and the distance between them and each data point is measured. Each data point is assigned to the closest seed, which becomes a cluster centroid, thereby creating an initial cluster assignment. Once this is completed the cluster centroids are recalculated by taking an average of all data points in each cluster. Data points are then reassigned if they become closer to an alternative cluster centroid. This should result in an improvement to the overall clustering solution as the distances between data points and cluster centroids decrease, as measured by the sum of squared deviations (Aldenderfer and Blashfield, 1984). This process is then repeated until a convergence criterion is met (Gale and Longley, 2012), where no further data points move between clusters and the variability within clusters has been minimised.

The design of the algorithm results in k-means having a constant weight function, where all data points belonging to a cluster have an equal influence on the centroid of that cluster; meaning that all clusters are created equally. The end results of the process are optimum clusters (for that initial seed assignment) where they will contain objects that are as similar to each other as possible, and individual clusters will be as dissimilar to each other as possible. Once the process is complete, the cluster means can be examined for each variable, allowing the distinctiveness of the clusters to be assessed and descriptions formed (Everitt et al., 2011). The comparatively simplistic nature of k-means makes it one of the most commonly used methods in geodemographics (Harris et al., 2005). The algorithm is however sensitive to the initial assignment of cluster seed points. This may result in a single completed iteration of algorithm not achieving an optimum solution. In addition, traditionally the k-means algorithm is computationally expensive for large datasets (Osamor et al., 2012). The algorithm can take hours or days to complete depending on the size of dataset and the processing power available.

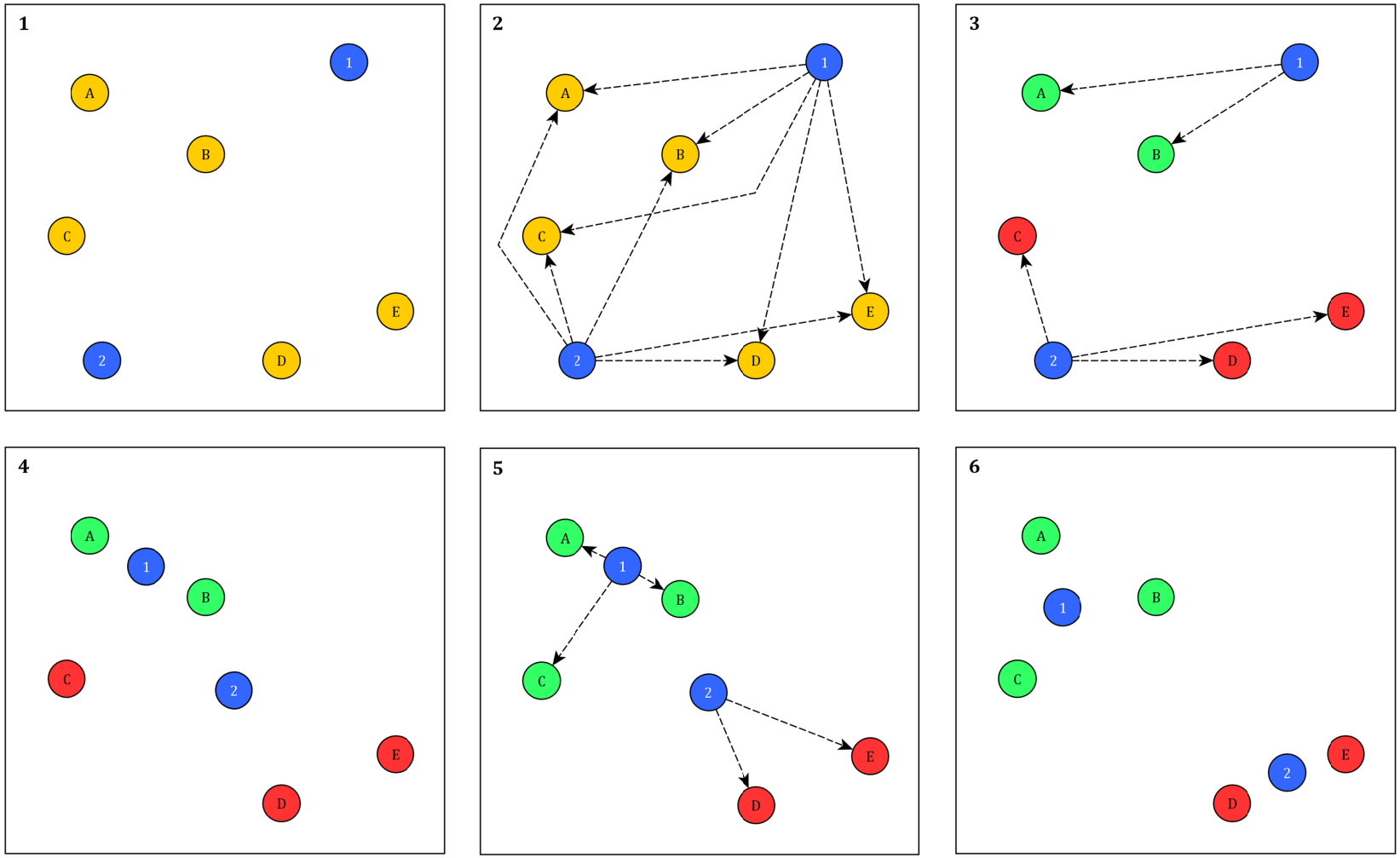


Figure 6.4: The k-means clustering process

The k-means algorithm was used to create the 2001 OAC (Vickers and Rees, 2007). As the aim for the 2001 OAC was to be a hierarchical classification, the algorithm was adapted to accomplish this task. Two methods were proposed; a top down approach and a bottom up approach (Vickers et al., 2005). The top down approach works by firstly clustering the entire dataset into n^1 clusters (with n being a pre-selected number), forming the top level of the hierarchy. The middle level of the hierarchy is produced by running the k-means algorithm on each of the n^1 datasets to produce n^2 clusters. Finally, the bottom level is created by using the k-means algorithm on each of the n^2 datasets to create n^3 clusters. The bottom up approach reverses this process and creates the bottom level of the hierarchy first, then the middle level and finally the top level. The top down approach was selected as it was believed to be "fundamentally better" and because "the [top] level was seen as the most important" (Vickers et al., 2005, p. 41–42). The top down approach was furthermore considered by Vickers et al. (2005) to be better because it meant that objects, in this case OAs and SAs, were always clustered, as opposed to cluster centroids (as is the case in the bottom up approach). Vickers et al. (2005) explain the negative impact this could have on a final clustering solution. Cluster centroids will rarely be representative of an entire cluster. An object that has little resemblance to the cluster centroid can still be included within that particular cluster. Going up a hierarchical level and clustering using these centroids can create clusters containing objects with little in common, and thus result in clusters with low homogeneity.

6.8.1.2. Ward's hierarchical clustering algorithm

Ward's hierarchical clustering algorithm seeks to cluster a large array of objects into smaller, mutually exclusive groups (Ward, 1963). The composition of these groups will consist of objects that are as similar to each other as possible. Hierarchical clustering is an agglomerative, or bottom up, approach to clustering. All objects are initially separate and the algorithm searches all possible pair combinations, selecting the pair that minimises the within cluster variance (Szekely and Rizzo, 2005). This process continues until all objects have been merged to form one single cluster. As the algorithm relies on minimising the within cluster variance, clusters are not guaranteed to be optimal. The minimised within cluster variance is akin to giving the shortest route priority. In some instances this can mean that the most appropriate route does not get taken, leading to reductions in individual cluster homogeneity. To illustrate the relationships between objects a dendrogram can be produced. This forms a tree-like diagram expressing the links between objects. All objects are joined together at the base, with branches linking

objects that are clustered together, leading to the formation of new nodes. The length of each branch, the cophenetic distance, indicates the strength of the relationship between those particular objects or nodes.

Ward's algorithm was considered for use in the 2001 OAC (Vickers et al., 2005). The advantage of using this method was that as the 2001 OAC was required to be hierarchical, there would only be a need to run the algorithm once and subsequently decide on the three levels (as depicted on a dendrogram) where the clusters would be. The method was however rejected due to its unsuitability for such a large dataset. Ward's algorithm can only be run on small datasets of approximately 1,000 objects or fewer (Vickers et al., 2005). In an attempt to resolve this, Vickers et al. (2005) first clustered the data using k-means into 1,000 objects and then applied Ward's algorithm. The sizes of the clusters produced using this combined method ranged from containing 125,000 OAs to 3. Detailed reasons on why this occurred are provided in Vickers et al. (2005), but utilisation of Ward's algorithm to create the 2001 OAC would have led to inconsistently sized clusters which revealed little about the UK.

6.8.1.3. Partitioning Around Medoids

The Partitioning Around Medoids or PAM algorithm (Kaufman and Rousseeuw, 2005) is an alternative form of partitional clustering. PAM acts to find a representative object whose average dissimilarity to other data points is minimal, called a medoid, for each cluster. The algorithm selects an object as a medoid for each of the predefined number of clusters. Those objects not initially selected are grouped with the medoid that shares the most similar characteristics. The medoids are then swapped for another object and the process is repeated until all objects have been assigned to medoids and the optimum clusters are produced. The objects are selected at random, and are actual data points assigned to clusters based on their closeness in the matrix (as opposed to points within Euclidean space which is used for algorithms like k-means). PAM is less sensitive to outliers because the process of assigning medoids uses a median rather than the mean in the optimisation procedures. The PAM algorithm can be more robust than k-means because it seeks to minimise the sum of dissimilarities rather than the sum of squared Euclidean distances, although this will vary depending on the type of clustering desired. Due to nature of the algorithm, PAM can be computationally expensive as each object is compared with the entire dataset (Ng and Han, 1994). It can also produce singleton

clusters when atypical objects are used in combination with a relatively small cluster number assignment (Kaufman and Rousseeuw, 2005).

6.8.1.4. Consensus clustering

Consensus clustering is an alternative method to contemporary clustering methods that rely on a single algorithm to find an optimal clustering solution. Instead it uses a combination of traditional cluster methods to produce more consistent results. This method was first proposed by Monti et al. (2003) and can be advantageous as through the combination of multiple clustering methods, the weakness of one particular algorithm can be offset with the advantages of another. It allows common traits between different clustering methods to be identified (where different algorithms identify the same clusters), while also revealing any differences between them. This method of clustering will subsequently seek to find a consensus which best represents the dataset (Goder and Filkov, 2008). A number of metrics can be used to indicate which clustering method provides the optimum outcome, and what the number of inherent groups within the data are (Simpson et al., 2010). This is particularly helpful when using methods such as k-means, which rely on random seeding to allocate initial clusters (Monti et al., 2003). Some of the benefits of consensus clustering include the ability to generate better clusters whilst being sensitive to noise and outliers in datasets (Nguyen and Caruana, 2007). The optimum result of consensus clustering is one where there is little difference between methods used as there can be more confidence in the clusters formed if more than one method has identified them in the data.

Geodemographic classifications with methodologies open to scrutiny have not previously utilised consensus clustering. The primary reason, other than its relative newness, is the issue of utilising the method on the large datasets which are commonplace with geodemographics. The current computational power available would not allow a geodemographic classification to be created from consensus clustering at a national scale at the finest resolution on a standard desktop workstation (Gale and Longley, 2012). Current computational efficiency means that computing all permutations for all clustering algorithms would take significantly longer than alternative methods. This problem currently prohibits a wide adoption of the method. As the 2011 OAC is designed to have a methodology that is reproducible, it would not be suitable to utilise a clustering method which most users would not be able to readily use. An additional issue associated with consensus clustering is that the metrics produced to

help identify the optimum algorithm or numbers of clusters rely on a definition of 'optimum' that can be quantified, something that is often only possible to do so qualitatively. Consensus clustering may lead to the best and most robust clusters from a statistical point-of-view, but they may be inherently poor at representing the socio and demographic variations across the coverage of a geodemographic classification.

6.8.1.5. Other clustering techniques

There is a plethora of clustering techniques available to use, although not all methods have been utilised historically for geodemographical purposes. Arabie et al. (1996), Gordon (1999) and Xu and Wunsch (2009) provide a more thorough review of clustering techniques.

AMOEBA (A Multidirectional Optimal Ecotope-Based Algorithm) is a procedure developed by Aldstadt and Getis (2006). It assesses the spatial association of a mapped object to surrounding objects and identified hot and cold spots (Jankowska et al., 2008). Whilst it has been designed to handle geospatial data as it searches for spatial association at the finest scale, it is not appropriate for use in the 2011 OAC. The way in which AMOEBA identifies hot and cold spots puts more emphasis on the statistical significance of the resulting clusters than it does on assigning all objects to clusters. There is a possibility that not all objects will have a cluster assignment, something which is a requirement of the 2011 OAC. Jankowska et al. (2008) do, however, introduce the possibility of modifying the method to address this issue.

The concept of fuzzy classification has been discussed briefly in Section 6.2 within the context of uncertainty within a final classification. Fuzzy classification can also be used to form the basis of a classification. Fuzzy classifications group objects into fuzzy sets (Zadeh, 1965). For geodemographic classifications this can be visualised as assigning all objects different shades of grey, rather than definitively black or white (either a member of a cluster or not). Fuzzy classification is a product of the belief that large datasets are inherently too complex to be grouped into well-defined clusters, and to do so is an oversimplification (Gordon, 1999). This can be said to be truer for objects that are on the edge of clusters as these will be increasingly dissimilar to those found in the centre of the cluster (Vickers et al., 2005), and are in fact likely to share more characteristics with objects in other clusters. It can therefore be said that the propensity of an object to share characteristics of objects found at the centre of its cluster reduces the further away it is

from that cluster's centroid. A fuzzy classification system uses this property to classify each object as having a proportional membership to all the final clusters, rather than belonging to only one (Voas and Williamson, 2001; Everitt et al., 2011). A fuzzy method also allows for geographic inter-cluster differences to be identified. On a national level, a cluster may be superficially similar, but at a local scale there may be variations between different areas. An example of this is the 'Countryside' cluster from the 2001 OAC. Across the UK 12.4% of OAs are assigned to this group, but the composition of the group is different in Scotland when compared to the rest of the UK. This difference is not identifiable from solely looking at the cluster assignment, and requires a more detailed analysis. These differences within cluster groups that cover large geographical areas exist because geographical phenomena are influenced by global effects first and local effects second (Bailey and Gatrell, 1995; Estivill-Castro and Lee, 2000). Fuzzy classification is one method that can handle this inter-cluster variation. Fuzzy versions of traditional clustering methods have been developed, such as fuzzy c-means (Bezdek, 1981). Feng and Flowerdew (1998) provide a description of how a fuzzy classification can be created.

Whilst a fuzzy classification is more likely to provide a more realistic presentation of reality (Vickers, 2006), such a classification would be harder to interpret. The importance of a classification simplifying a complex dataset into an easily understandable format meant that this type of fuzzy methodology was not used for the 2011 OAC. Although, similar to the discussion in Section 6.2, the uncertainties created by the fuzzy aspects of the 2011 OAC were not completely discarded. Including them as an ancillary part of the creation process will aid the creation of similar tools presented by Slingsby et al. (2011).

Vickers (2006) investigated the use of artificial neural networks as a method of clustering for the 2001 OAC. A particular technique, known as a Self Organising Map (Kohonen, 1984) was used to create one of the first open geodemographic systems and a predecessor of the 2001 OAC, the 'GB profiles' geodemographic system (Openshaw, 1994). GB profiles clustered Enumeration Districts from the 1991 UK Census, and as the Self Organising Map technique is an unsupervised algorithm, it was able to identify an ideal number of clusters without having them pre-specified (Kohonen, 1998). There are however, problems with other aspects of methods which utilise the artificial neural network framework which make it unsuitable for the 2011 OAC. Vickers (2006) discounted their use in the 2001 OAC because the method has hidden layers that make

understanding the assignment process difficult. Although the algorithms that underpin the method are open, the level of difficulty associated with understanding the technique makes it unsuitable for adoption with the 2011 OAC; especially as one of the fundamental principles of the classification is to have an open, understandable and fully documented methodology.

An important factor when considering a suitable clustering method is the ability to understand the processes behind the algorithm. The process can therefore be documented and can be subsequently used to aid interpretation of the final cluster output by users. The importance of understanding how a clustering algorithm works was deemed a key component of creating the 2011 OAC and considered a key criteria when selecting which method to use.

6.8.2. Distance measures

The primary purpose of clustering algorithms is to minimise variability within cluster objects and maximise variability between cluster objects (i.e. distances between objects are minimised within clusters and maximised between clusters). This is calculated based on the mathematical distances between all objects, which can relate either to actual distances or arbitrary values in multidimensional space (Johnson and Wichern, 2007). Distances can be measured in terms of similarity, where lower values signify that objects are more similar, or dissimilarity, where larger values signify that objects are more similar. There are no definitive rules on which type of measure should be used, but in most cases knowledge of the data being clustered should be utilised to provide guidance on the most appropriate measure to use. For example, a measure may be more suited to a certain data type, such as continuous, discrete or categorical. This knowledge of the data is particularly important as the majority of standard distance measures assume that the clustered objects are continuous in nature (Anderberg, 1973). There are a number of different measures as outlined by Everitt et al. (2011), but only a few are commonly used and these themselves share similar attributes. Below is a brief summary of a selection of distance measures:

i) **Euclidean distance**: The geometric distance in multidimensional space between objects. It is calculated by taking the square root of the sum of squares of the differences between the values of objects. This is one of the more popular distance measures due to its relative simplicity. However, despite reaching the

maximum number of iterations allowed within a clustering algorithm it does not always converge.

ii) **Squared Euclidean distance**: This is the same as Euclidean distance but without the square root being taken and the distance values being squared. As it does not take the square root it gives a clustering algorithm increased performance, when compared to the Euclidean distance, due to better computational efficiency (Mouffron et al., 2008). The measure is better at handling large numbers of objects (Everitt et al., 2011) and helps convergence of a clustering algorithm in large datasets. It also acts to place an increasingly greater weight on objects that are further apart. However, it can emphasize any outliers since the square function magnifies larger values. As such, only a small number of outliers would be required to have an adverse impact on clustering solutions.

iii) **Manhattan distance**: This is the average difference across dimensions. In most cases, this distance measure produces similar results to the Euclidean distance. A key difference, however is that the effect of outliers (large differences between objects) is reduced since these values are not squared.

iv) **Chebyshev distance**: This measure can be used to identify two objects as different. It is calculated by taking the maximum distance between objects.

The selection of the type of measure to use for the 2011 OAC was dependent on the clustering algorithm used. Vickers (2006) states that for partitional clustering methods, such as k-means, there is no difference between using the Euclidean distance or the squared Euclidean distance. However, as Everitt et al. (2011) state, squared Euclidean distance has the capacity to handle large datasets better. There is however a difference in cluster outputs when using either of these measures with a hierarchical clustering method (Everitt et al., 2011). Vickers (2006) concludes that Euclidean distance is the preferred option for hierarchical clustering, and squared Euclidean distance for partitional clustering. While the other distance measures discussed could have been used for the 2011 OAC, it was deemed unnecessary to test their appropriateness for the new classification. This was because the distance measures compatible with the 2011 OAC dataset would likely have produced comparable results due to the similarities that the measures share.

6.8.3. Selecting a clustering method

Unlike the data preparation stage, it was felt that pre-defining the clustering method to be used on the final variable selection was appropriate. Any clustering method could have been used with the 2011 OAC, but the final choice was guided by the responses to the 2011 OAC user engagement. Respondents indicated a preference for the top-down hierarchical structure of the 2001 OAC. As a result, the use of a partitional clustering method and recreation of a similar structure with the 2001 OAC was selected. Different methods were considered, but ultimately using the same partitional clustering method and distance measure used on the 2001 OAC was judged to be the best way of ensuring that a similar cluster structure was created for the 2011 OAC. As such, k-means using the squared Euclidean distance measure was selected as the clustering method.

An advantage of using k-means as the clustering algorithm for the 2011 OAC meant that improvements made in the methodology, namely testing multiple data preparation procedures, could be subjectively assessed to see whether they led to the creation of a better classification than the 2001 OAC. Maintaining the same clustering procedure meant that any uncertainty, which would have been created using an alternative method, was removed. The selection of k-means method did however mean that the bias and generalisation associated with the reliance on global parameters, such as the number of clusters, needed to be accounted for (Estivill-Castro and Lee, 2000). To resolve this problem for the 2011 OAC, different numbers of clusters were tested using multiple iterations of the k-means clustering algorithm. Due to the random nature of the initial seed assignment of k-means, each iteration of the algorithm can produce different clustering solutions. Running multiple iterations of the algorithm on the same dataset for the desired number of clusters meant that an optimum solution was identified (see Section 6.8.6).

6.8.4. Cluster numbers and classification structure

Geographical phenomena are primarily the result of global order and secondly by local order effects (Bailey and Gatrell, 1995). Clusters created from the 2011 UK Census are therefore the product of both effects at the national (global) and the sub-national or regional (local) geographic scale. The relationship between global and local effects is not consistent, therefore the number of clusters directly impacts upon how these different effects manifest themselves in the final groupings. Deciding upon the optimum number of clusters in a classification is dependent on a number of factors. Considerations such as

the structure of the classification have a direct impact on the final clusters produced. A hierarchical classification, with a tiered structure, may require a different number of clusters to a classification that is non-hierarchical to best represent the characteristics of the resident population. For example, a hierarchical classification could have fewer clusters for its top tier that have full geographical coverage and provide more generic population summaries. The lower tiers could then offer more specific descriptions, albeit without full geographical coverage. In contrast, a non-hierarchical classification is more likely to include more clusters for its single tier in order to produce summaries of the population with adequate detail.

As the 2001 OAC's three-tiered hierarchical approach and cluster number, discussed in Sections 2.6 and 6.8.3, was deemed satisfactory by the 2011 OAC user engagement (see Section 4.3.2) there was an expectation that the 2011 OAC would be similar. There was not, however, an expectation of having identical cluster numbers to the 2001 OAC, in the same way there was no expectation to use the same variables. The 2011 OAC therefore aimed to have similar, but not identical numbers in a three-tiered structure with the retention of the Supergroup, Group and Subgroup terminology. The advantage of retaining a consistency in clustering methods between the 2001 OAC and the 2011 OAC was in maintaining a similar structure between the two classifications. Using a top-down clustering method meant that Subgroups still nested within Groups, and Groups within Supergroups. This can be seen as a desirable feature of the 2011 OAC as it makes adapting to this new version easier for those experienced in using the 2001 OAC.

As discussed in Section 2.5, there is no consensus on how to structure or form a geodemographic classification. Singleton and Spielman (2013) did however conclude that the hierarchical structure tends to be of three levels for UK classifications, albeit with different number of clusters per level. With no common methodological approach, the number of clusters required to best represent a population is unique to every classification. The alternative method for determining the number of clusters and hierarchy to use with the 2001 OAC came from personal correspondence with Professor Martin Callingham (Vickers, 2006). This led to the ideal number of clusters for each hierarchal level, and what their intended uses could be. The reasons given for the number of clusters used in the 2001 OAC and structure are summarised below:

- **Supergroups:** The highest level of aggregation should have around 6 groups. This allows for descriptive names and can be visualised easily.

- **Groups:** The middle level of aggregation should have around 20 groups. This would be useful for customer profiling and allows market propensity measures to be established with comparatively small surveys. Ideally the groups would have descriptive names.
- **Subgroups:** The lowest level of aggregation should have around 50 groups. This can be used for market propensity measures from the larger commercial surveys and government surveys. The groups do not need descriptive names.

The view that the clusters which form the lowest level of aggregation do not need names is contradicted by anecdotal evidence from users of the 2001 OAC who have suggested that having names at this level would have improved the classification. It should also be noted that commercial systems, such as Mosaic and Acorn name the clusters in their respective lowest aggregation level (Experian, 2010; CACI, 2013b). As such, the 2011 OAC can only be considered an alternative to these products if all levels of the classification hierarchy are named.

The final number of clusters used with the 2011 OAC was decided by utilisation of both quantitative and qualitative methods. The steps outlined in Section 6.2 by Everitt et al. (2011) provide some generic guidance on the approaches taken. Vickers et al. (2005) details the three main considerations in deciding upon cluster numbers in a classification:

1. The average distance from the cluster centre for each cluster number option should be the smallest possible.
2. The size of the clusters should be as similar to each other as possible.
3. The number of clusters should be as close to the perceived ideal as possible.

The first two issues can be quantitatively measured to determine how many clusters are needed to find the optimum solution. The third issue is more subjective, and has the most scope to vary between different classifications. It was therefore important to decide what an ideal number of clusters should be for the 2011 OAC. This would ideally be the maximum number of clusters that distinguish different characteristics of the UK population, which have similar number of OAs and SAs assigned to them.

Finding the optimum solution for the 2011 OAC was achieved through the methodical process of testing numerous outputs and cluster numbers. Utilisation of the k-means

algorithm produced a similar structure to the 2001 OAC forming a classification constructed in a top-down format. A final decision made on both the dataset and cluster numbers used to create the final 2011 OAC hierarchy was based on both qualitative and quantitative assessment. Decisions were based on testing multiple permutations of cluster numbers on the different datasets. The top level of the hierarchy had 5 to 9 clusters tested, while the middle and bottom levels had 2 to 4. These numbers were chosen as they resulted in a similar hierarchical structure to the 2001 OAC, without requiring an identical number of clusters. The dataset selected created the optimum clusters in terms of the number of groups, and their reflection of the population and distribution characteristics.

6.8.5. Cluster names and descriptions

The final clusters for each level of the 2011 OAC hierarchy were given names and descriptions to aid the understanding of a cluster's characteristics. Vickers (2006) noted that names and descriptions of clusters can be contentious due to their potential to reinforce negative stereotypes. As you move down a classification hierarchy the clusters relate to decreasing proportions of the population. This can give the impression that names for some of the smaller clusters relate more to individuals than the general area, risking invoking an ecological fallacy (Robinson, 1950). The extent to which this impacts the usefulness of a classification depends on the specific geographic area. A geodemographic classification is more useful when individuals can identify with the names and descriptions given to their local area.

To reduce the likelihood of individuals taking issue with names and descriptions, certain words and phrases were avoided where possible for the 2011 OAC. Words that made value judgments, unless they could be justified statistically, were avoided. Other words that were overtly negative or positive were also avoided. Where possible, language was used to imply that the characteristics of an area were a consequence of factors that have happened to the resident population and not because of them. This was done in keeping with the view that value judgements should be avoided.

The names and descriptions of clusters are based on their average characteristics. However, areas assigned to each cluster will differ in how much they conform to these average characteristics. As such, names could not be too specific, as these would correspond more significantly to the OAs and SAs closest to their assigned cluster's

centroid. Conversely, names that are too broad pose the risk of being too vague to offer any real insight into an area. In these circumstances the deviation found in the clustering process would be masked by similar names and descriptions being given for each cluster. As such, names and descriptions for the 2011 OAC were considered on a more simplistic scale of either bad or good for their intended purpose.

Cluster names can be considered as the user interface of any geodemographic classification, meaning that the underlying complexity of the cluster compositions has the potential to go unnoticed. Although other outputs from the clustering process are available to aid interpretation of the classification, it is still likely that the primary focus of many 2011 OAC users will be on the cluster assignment and their associated names and descriptions. Based on the issues identified in the naming process a set of guidelines on how the 2011 OAC clusters were named can be produced. The extent to which these guidelines have been kept depended on the final composition of each cluster created. They do however, offer pertinent advice for the naming of any geodemographic classification:

- All of the clusters created for each hierarchical level should be named to aid user interpretation, regardless of the use of the classification.
- The same names used by other geodemographic classifications cannot be used. Doing so would add confusion and suggest that the clusters in different classifications can be compared – this is especially true with the 2001 OAC and 2011 OAC. While the 2011 OAC has been designed to be structurally similar to the 2001 OAC, the clusters themselves are unique and share no links.
- Neutral phrasing of the names and descriptions is important to improve the likelihood that the clusters will be accepted as providing a realistic interpretation of areas.
- The names should avoid being too specific but still provide unique descriptors of each cluster.

6.8.6. Optimising clustering algorithm iterations

The use of k-means as the clustering algorithm for the 2011 OAC meant that several steps were taken to ensure that optimum outputs were produced. As previously discussed, the random nature of the k-means algorithm and its initial seed assignment means that each iteration of the algorithm can create different clustering solutions. As such, it was

advantageous to run the algorithm multiple times in order to find the optimum solution. To achieve this, the mean WCSS value (as discussed in Section 6.6.3) for each iteration was used, with the iteration that had the lowest WCSS value considered to be the optimum result. Singleton and Longley (2009b) suggest that the optimum number of iterations for k-means is 10,000; however, this figure should not be used indiscriminately without first considering the size of the dataset that requires clustering. For example, the 2001 OAC, with its 223,060 OAs and 41 variables creating 9,145,460 data points, takes over one day to run through 10,000 iterations of k-means (assuming the use of a high specification desktop PC). As such, the performance requirements can be a prominent factor into deciding how many iterations of k-means are performed, rather than using an arbitrary value. Estivill-Castro and Lee (2000) noted that the GIS community seemed more focused on producing optimum clusters rather than the computational efficiency. The focus for the 2011 OAC was therefore to create optimum clusters in the most efficient way possible.

To ascertain the optimum number of iterations for the k-means algorithm for the 2011 OAC, eight subsets of the 2001 OAC dataset were clustered, with each subset representing a differently sized geographic area. This was done to test the hypothesis that the size of a dataset impacts the efficiency of the clustering process and the optimum number of algorithm iterations. The results of this are presented in Figure 6.5 where these eight differently sized areas, ranging from the ward of Bloomsbury in London to the whole UK have been clustered and the resulting WCSS values for each iteration are recorded. It is clear that the size of a dataset has no bearing on how the k-means algorithm functions, with all eight datasets being equally unstable from a clustering prospective. Across the eight areas, each iteration produced a unique WCSS value. The only exception was in Bloomsbury, where the smallest WCSS value was achieved 94 times and 47% of values were not unique. This indicates that 10,000 iterations may not have been enough to ensure an optimum clustering solution, and smaller WCSS values may have been achieved if more iterations had been performed.

Although the clustering solution that uses the smallest WCSS value can be considered optimum from a statistical perspective, it does not guarantee the robustness of any outputs. A global measure like the WCSS value is explicitly non-spatial, and as such cannot guarantee that any spatial output represents geographical phenomenon, such as neighbourhoods, optimally.

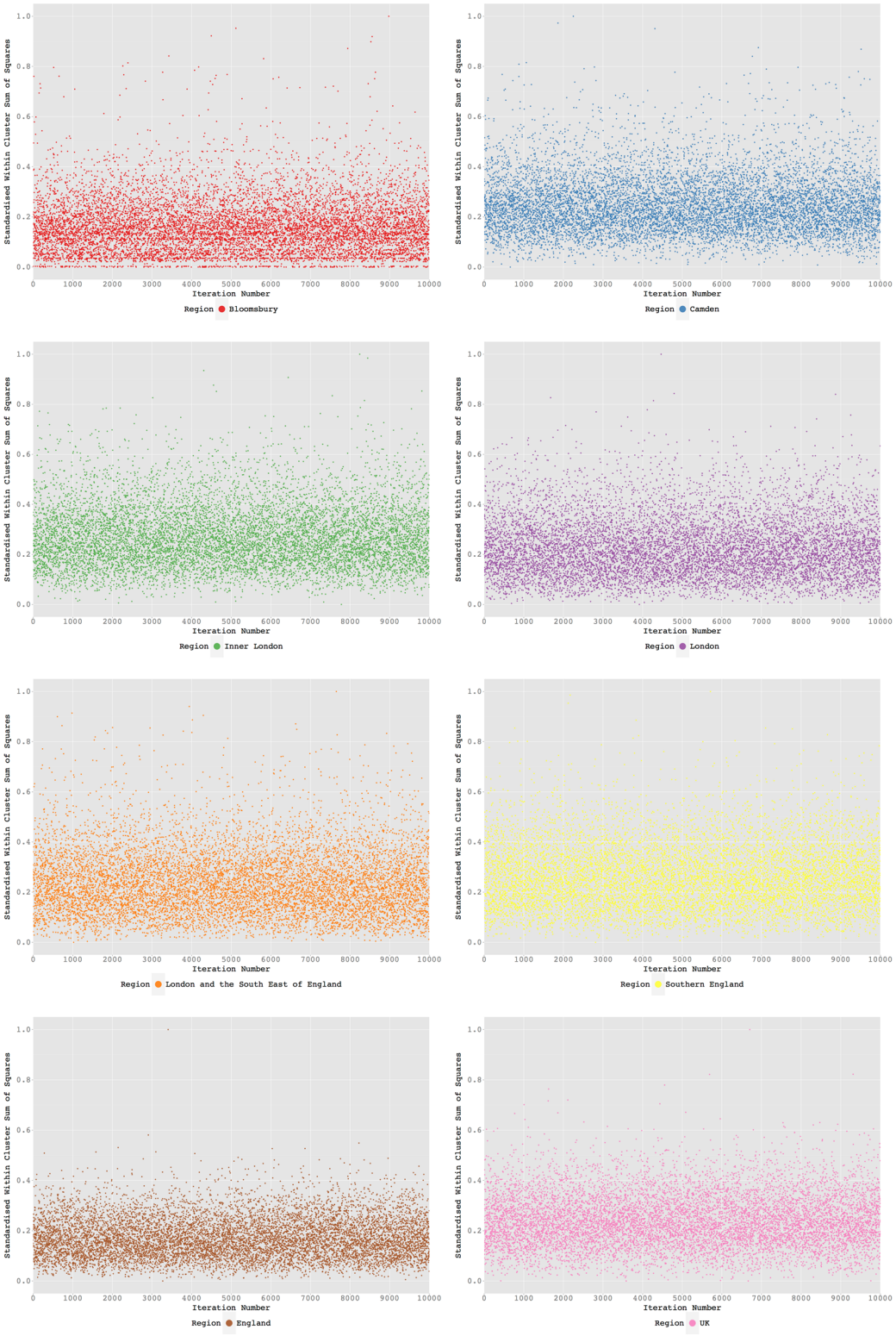


Figure 6.5: WCSS values for n k-means runs of the 2001 OAC

The implications for the 2011 OAC dataset are that in theory, *n* iterations of k-means could be performed and an optimum cluster solution never achieved. The likelihood of finding an optimum solution are increased by running the algorithm more times, but computationally, a cut-off point needs to exist, and the 10,000 value stated Singleton and Longley (2009b) is appropriate for this reason. This multi-iteration approach is the most appropriate method when utilising the k-means algorithm, but if the same dataset needs to be clustered multiple times and be able to produce the same results (for example reproducing the 2011 OAC for further analysis and critique) it can be unreliable.

A solution to this problem would be to fix the initial seed points. This would use the seed points from the best of the 10,000 previous iterations, meaning that only a single iteration of the algorithm would be required to produce an optimum clustering solution. Although this method would make future re-clustering of the 2011 OAC dataset a more efficient process, it would compromise its use with any other dataset. This is because the ideal seeding locations change for every dataset. Fixing the seeding points for optimum use with the 2011 OAC dataset would therefore lead to non-optimum clustering solutions with any other dataset. This would have a significant negative impact on the ability for the 2011 OAC's methodology to be more widely adopted. The multi-iteration approach was therefore the best overall option as it allows the optimum clustering solution (within the 10,000 iteration computation limit) to be found for every dataset that uses it, while being as efficient as possible.

Although the clustering solution that uses the smallest WCSS value can be considered optimum from a global statistical perspective, this criterion alone is not necessarily the sole or even the most important aspect of a finalised classification. A global measure like the WCSS value is non-spatial, and as such cannot guarantee that any spatial output fully accommodates geographical considerations optimally, such as neighbourhoods or the size and configuration of areal units. Mandelbrot (1982b) argued that good models, or in this case clustering solutions, which generate spatial predications that can be mapped or visualised should 'look right'. In the context of a geodemographic classification this involves assessing whether the geographic distribution of clusters resemble what is known about an area. If the output of a clustering procedure with the lowest WCSS produces unrealistic geographic distributions of clusters, then it cannot be considered as the optimum solution. Mapping the best clustering solution is therefore vital before it can be declared 'optimum'. Despite the issues of clustering utilising the lowest WCSS value, the output from this should still be considered the optimum outcome. Ideally the

optimum clustering solution from a statistical perspective will also 'look right' when visualised. If this is not the case then it provides a good starting point to explore the clustering solutions considered non-optimum due to their WCSS values.

6.8.7. Optimised versus non-optimised clustering algorithm iterations

The need to use an optimal cluster output for the 2011 OAC was examined using the datasets and WCSS values shown in Figure 6.5. Using the clustering solution with the lowest WCSS value creates clusters that are as homogenous as possible, while the clusters themselves are more heterogeneous. Conversely, the highest WCSS value means that each individual cluster is more heterogeneous and differences between clusters are smaller. The impact of using a cluster solution produced using the lowest (or optimum) and highest (or non-optimum) WCSS values is shown for Camden in Figures 6.6 and 6.7 and for London and the South East in Figures 6.8 and 6.9. Although the designations of 'optimum' and 'non-optimum' are based on statistical outputs alone, a comparison between outputs derived from the non-spatial WCSS statistic provides an opportunity to assess the extent to which clustering solutions derived from these different values 'look right' (Mandelbrot, 1982b).

The results from both areas indicate that there is a difference between using an optimum and non-optimum clustering solution. In Camden there is a visual difference in the number of OAs that are assigned to 'Cluster 3' between the two solutions. Additionally, the distinction between 'Cluster 5' in the southern parts of Camden and 'Cluster 2' in northern Camden shown in the optimum cluster solution does not exist in the non-optimum version. The relative small size of Camden makes it difficult to judge which of the two solutions 'looks right'. The optimum solution offers more variation in terms of OA cluster assignment, but without knowing the micro-level dynamics of Camden it cannot be considered to be the better option. The most obvious difference between optimum and non-optimum solutions for London and the South East is the number of OAs assigned 'Cluster 2'. Although another notable difference is 'Cluster 3' being more dominant coupled with a reduction of 'Cluster 4' in London in the non-optimum clustering solution. The larger study area means that the optimum solution can be identified as the one which looks the most visually accurate. The greater variation in cluster assignment in London reflects the different population characteristics found, and the amount of variation in rural areas is greater than the non-optimum solution suggests.

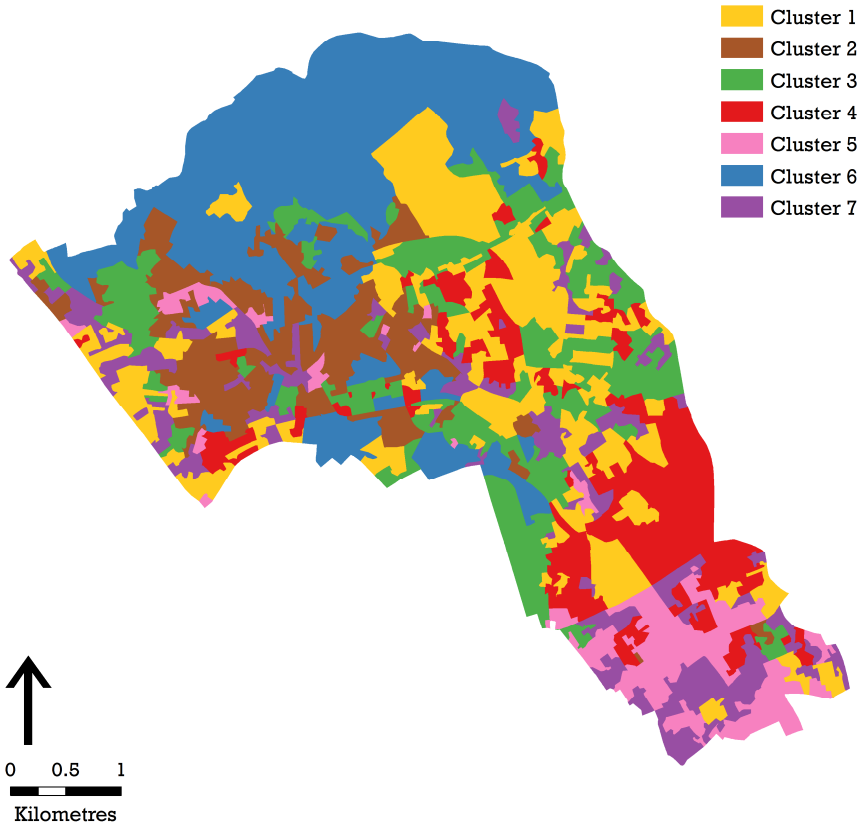


Figure 6.6: Cluster solution using lowest WCSS value for Camden

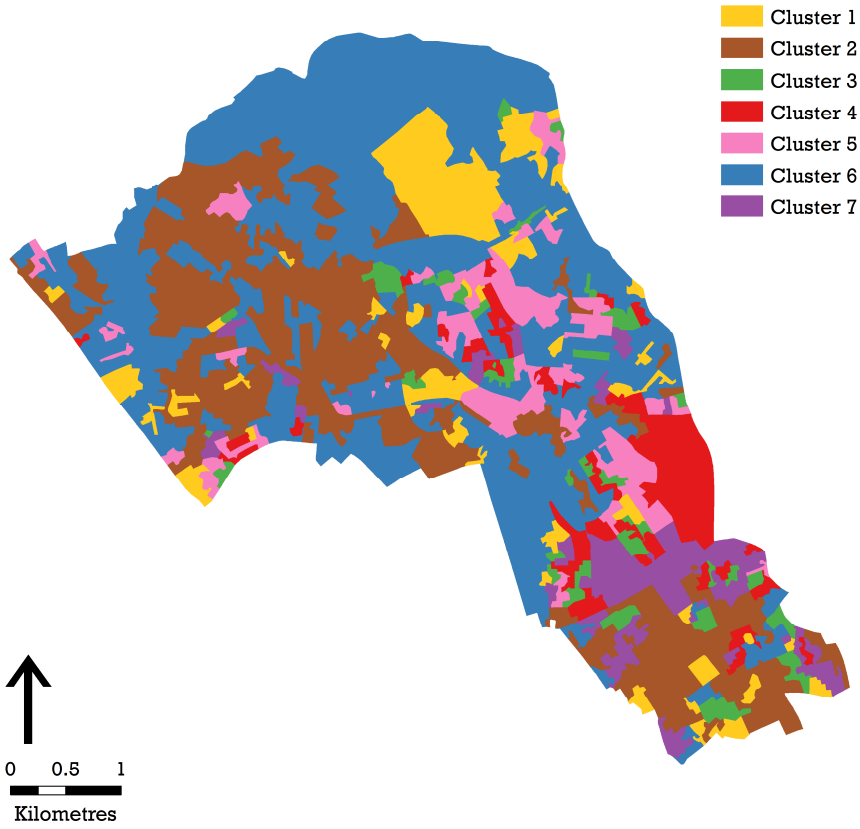


Figure 6.7: Cluster solution using highest WCSS value for Camden

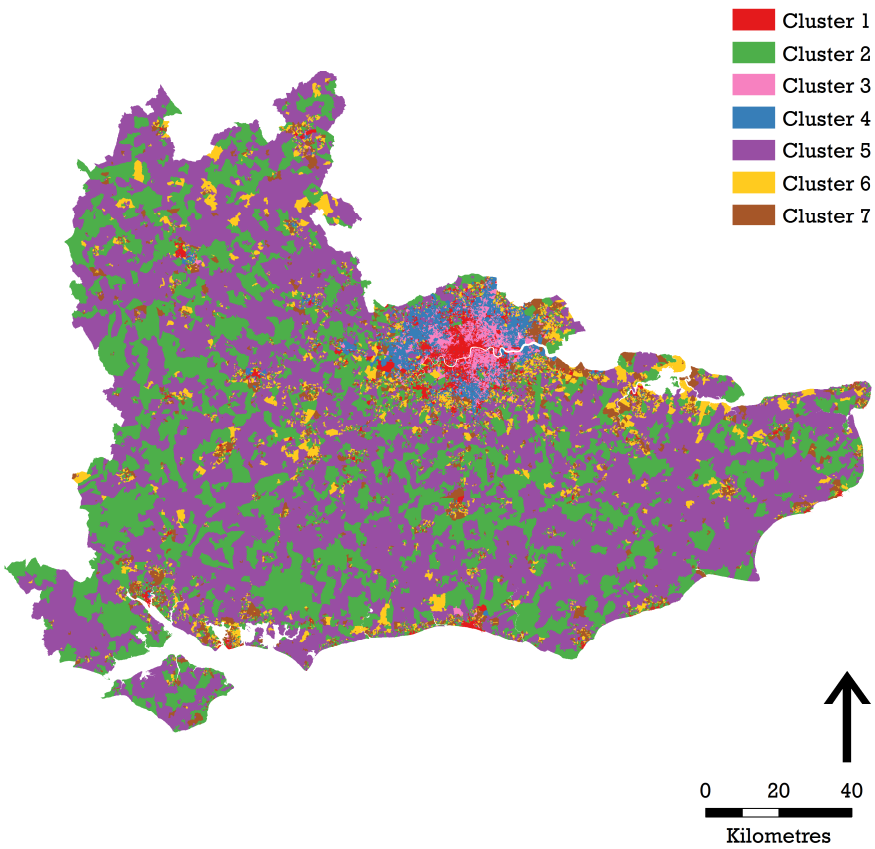


Figure 6.8: Cluster solution using lowest WCSS value for London and the South East of England

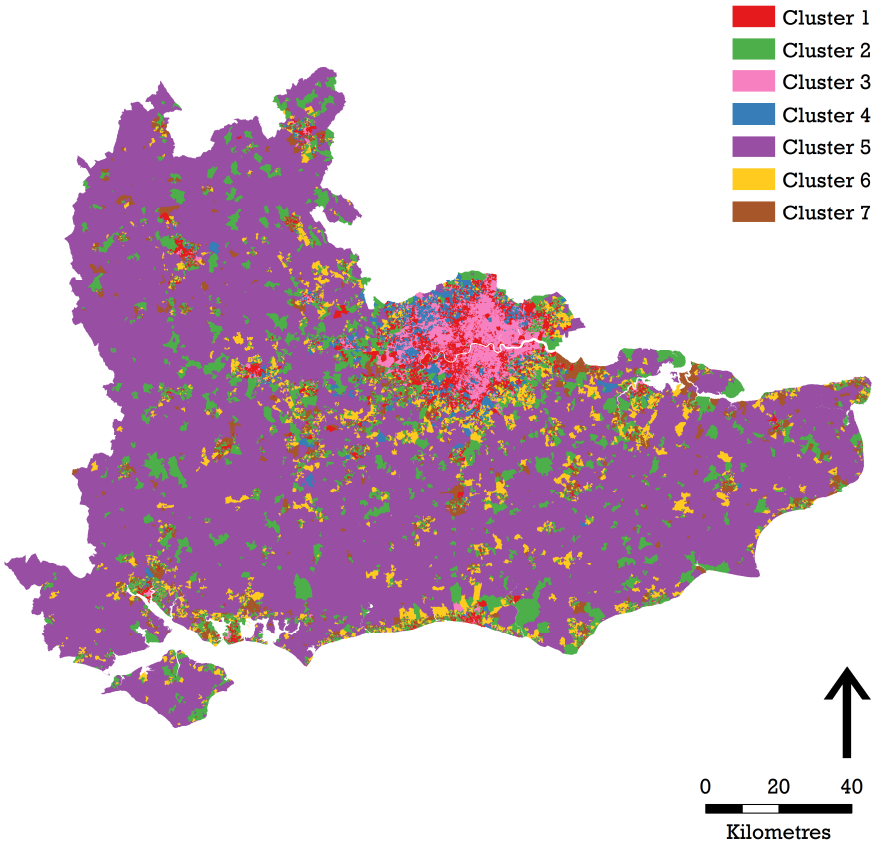


Figure 6.9: Cluster solution using highest WCSS value for London and the South East of England

The explanation of the significant differences between the optimum and non-optimum cluster solutions is more complex than the fact that individual clusters become less homogenous and differences between all clusters become smaller. Some clusters have a very different composition between the optimum and non-optimum cluster solutions, to the point where what visually looks like the same cluster actually represents different cohorts of the population. This is shown in Figure 6.10 – cluster profiles with the optimum and non-optimum cluster compositions overlaid on each other. The majority of clusters in Camden have fundamentally different compositions between the optimum and non-optimum solutions. This means that any similarities in the geographic distribution between the same clusters in these cases are only a coincidence. One exception to this is 'Cluster 4', where despite having different geographical distributions between the optimum and non-optimum solutions the cluster still represents similar population characteristics. The other exception, although to a lesser extent, is 'Cluster 2', but again this has different geographical distributions between the two cluster solutions.

The differences in cluster composition between the optimum and non-optimum solutions for London and the South East are less pronounced than those for Camden, but no cluster remains stable between the two. The smaller variations in cluster composition between optimum and non-optimum solutions for London and the South East compared to Camden may suggest that larger datasets (50,785 clustered OAs compared to 734) are less susceptible to change. This is however irrelevant for geodemographic classifications as any cluster compositions that differ from that offered by the lowest WCSS value are considered undesirable, unless they can be shown to offer a solution that is more representative of that area's characteristics than the more statistically robust option.

The differences between optimum and non-optimum clustering solutions created for Camden and London and the South East highlight the inherent random nature of k-means. The substantially different composition and geographic distribution of the cluster groups between these solutions highlights how sensitive k-means is to the initial random seeding of cluster sites. This is why steps, such as the 10,000 iteration value discussed in Section 6.8.6, were required to ensure an optimum outcome for the 2011 OAC dataset. A statistically optimum clustering solution identified using 1,000 iterations for example, may have corresponded to a statistically non-optimum solution when 10,000 iterations were used. The resulting differences in cluster composition between these two solutions may have been minimal, but these changes could still have a direct impact on how the final clusters and classification look.

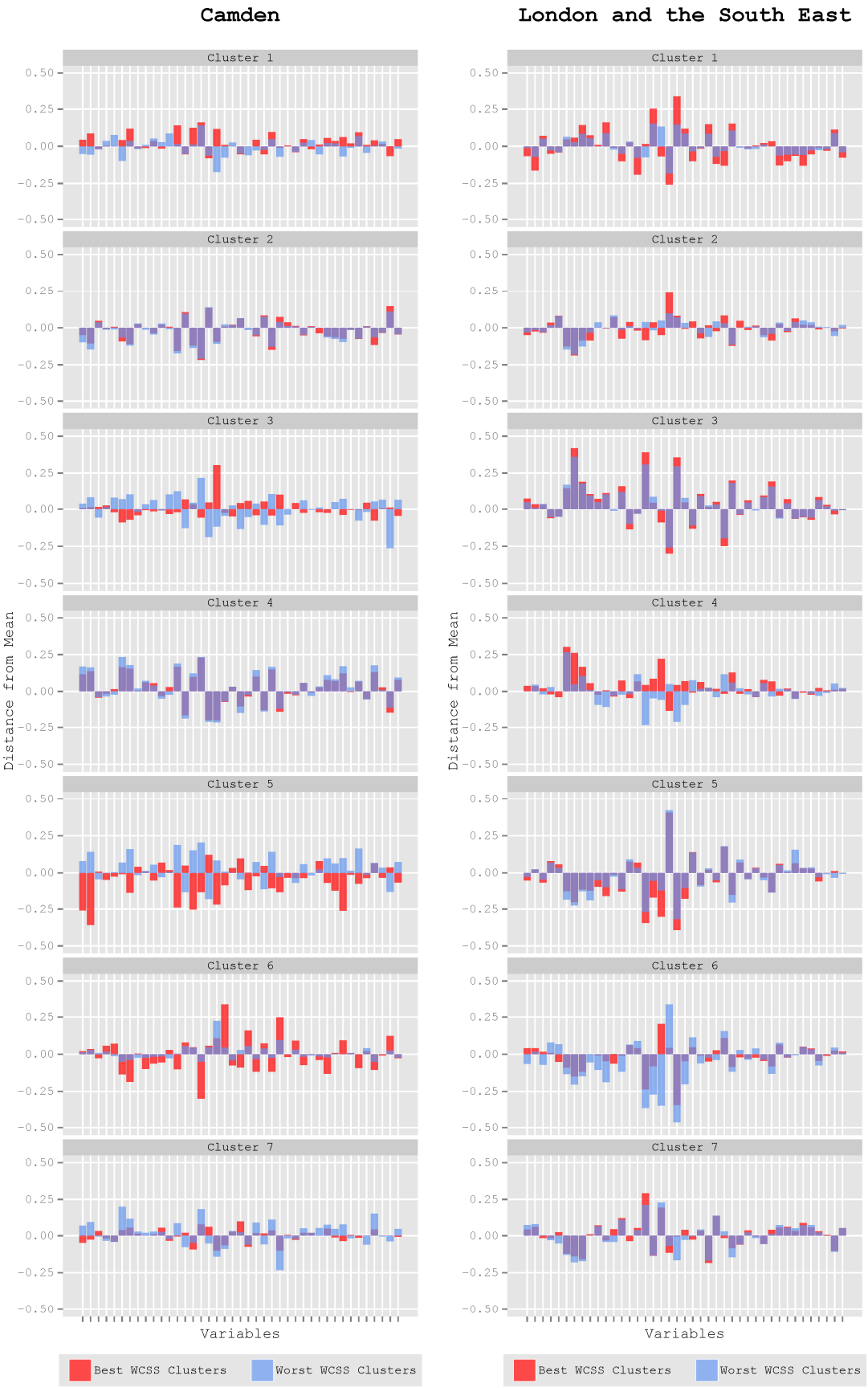


Figure 6.10: Cluster profiles using lowest and highest WCSS values

There needs to be confidence that the clusters of a geodemographic classification are as optimal as possible, both from a statistical perspective and in terms of what visually looks the most right. The London and the South East example indicates that there is a level of correlation between the two aspects of optimisation, where the utilisation of the lowest WCSS value results visually in the best cluster distribution. This is less clear in the Camden example, suggesting that while using the lowest WCSS is more likely to produce the best visual output, this becomes less certain when clustering a smaller set of objects. Overall, it can however be stated that clusters formed using the lowest WCSS value found from running 10,000 iterations of the k-means algorithm are likely to be both statistically and visually optimal.

6.8.8. Reproducible Clustering

An objective in the creation of the 2011 OAC methodology was to make it as reproducible as possible. Having the majority of the operations performed in the command line program R (R Development Core Team, 2011) helped to achieve this. This allows the scripts used in the creation of the 2011 OAC to be accessed by others, either in recreating the 2011 OAC or by modifying the methodology for their own particular uses. An important component of this is the ability to produce the same clustering results from the same set of variables each time. This consistency is vital to external evaluation and critique of the clustering process, as well as providing stability to an inherently unstable process.

As discussed in Sections 6.8.6 and 6.8.7, it is important that the k-means algorithm is run repeated times to ensure an optimum result is found. Using the lowest WCSS value after 10,000 iterations increases the chances of identical clusters being created every time the clustering process is run. What this will not do however, is order the clusters in the same way. The random initial assignment of seeds for k-means means that what was assigned as cluster 2 in one cluster run, could be assigned as cluster 4 in another. Therefore, whilst the composition and number of OAs assigned would be the same, they would be given different individual assignments. This is demonstrated in Figures 6.11 and 6.12, which shows the identification of the same clusters in Camden for different cluster runs, but with different designations.

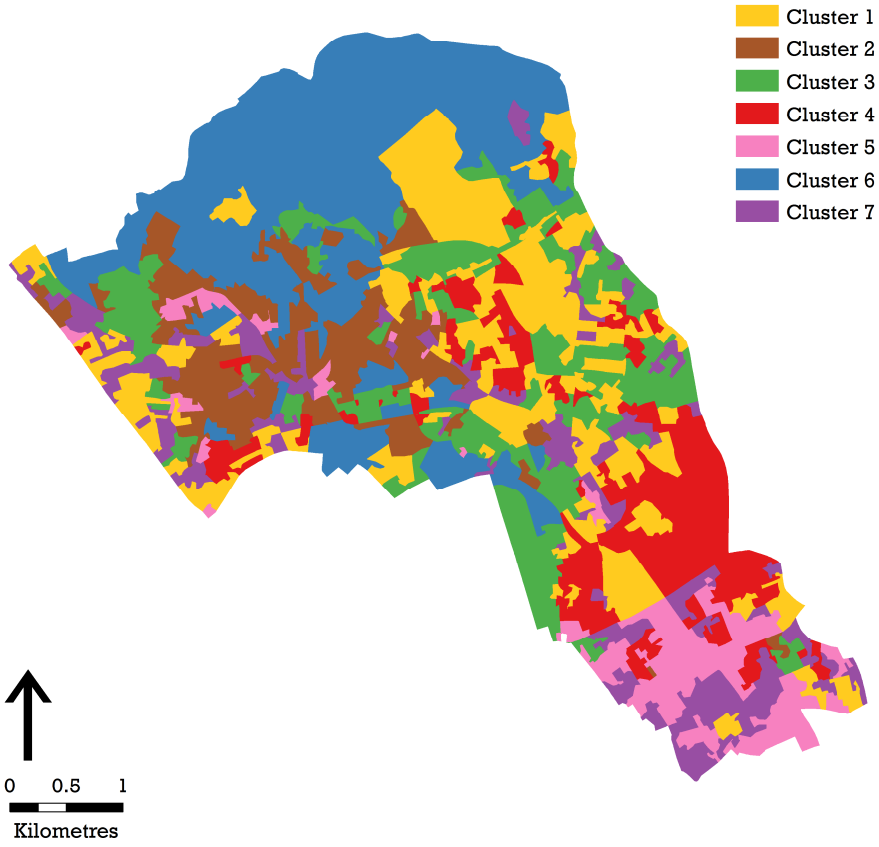


Figure 6.11: Cluster solution for Camden after first k-means run

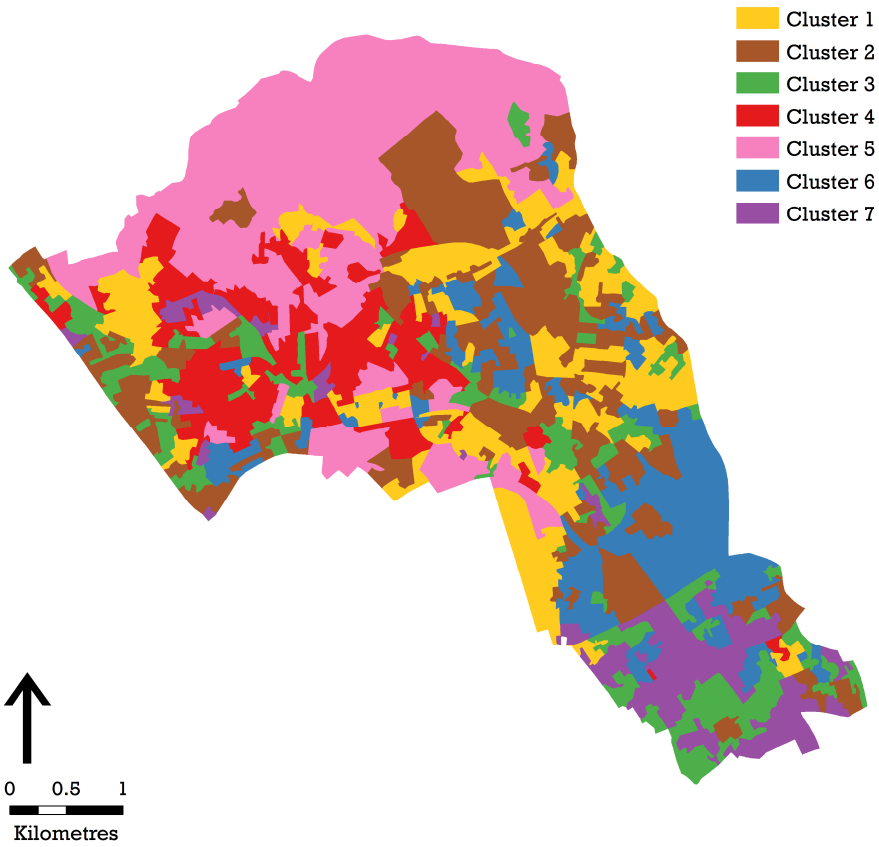


Figure 6.12: Cluster solution for Camden after second k-means run

Different designations for the same clusters are a natural consequence of using k-means, but are an undesirable feature for a fully reproducible clustering objective. A solution to this problem is to use each cluster's unique WCSS value created after the initial clustering process to re-order the clusters. Subsequent clustering that produces the same clusters but in a different order will also produce the same (or very similar) WCSS values. The WCSS values on a particular run can then be matched with those from the first run to re-order the clusters. This will give the appearance that each clustering run produces the same clusters in the same order each time.

The alternative method of ensuring the same clustering outputs every time the k-means algorithm is run is to fix the initial seed points. This method was discussed in Section 6.8.6 and discounted because it would mean that the code created for the 2011 OAC could only be used on that dataset. As this would have a negative impact on the adoption of 2011 OAC's methodology for use with future geodemographic classifications it was discounted.

6.9. Conclusions

This chapter has outlined the process that underpins the creation of the 2011 OAC. The methodological approach used in the creation of the 2011 OAC is designed to offer a robust and rigorous approach to a field where there is no predefined set of rules on how to produce a geodemographic classification. The decisions made during the creation of the 2011 OAC cannot all be quantitatively assessed; indeed the qualitative aspects in creating a geodemographic classification are perhaps the most important. While statistical and mathematical approaches can help inform a classifications' creator, subjective decisions need to be made on what will produce the best final outcome. In this context, the best can be defined as a geodemographic output which provides the most use to the greatest number of people, which is something that statistical and mathematical tests alone cannot guarantee.

Creating a geodemographic classification can lean heavily on past methodologies or create new and expansive methodologies utilising methods and procedures never previously applied in the field. Ultimately, whatever the techniques used, the aim is the same: to create a geodemographic classification which can summarise the varying characteristics of the study population. The methodology for 2011 OAC is an evolution of that used for the 2001 OAC. This is a reflection of the views expressed by current and

past users of the 2001 OAC that they favour the structure, outputs and general characteristics of the classification. As such, improving the 2001 OAC methodology, rather than creating something new, has been the focus.

This focus has meant identifying areas of the 2001 OAC methodology that could be improved to create a better classification, and other areas where modifications were less likely to fundamentally alter the 2011 OAC. Based upon the different stages of cluster analysis as identified by Milligan (1996) and expanded by Everitt et al. (2011), decisions were made to test multiple rate calculation, transformation and standardisation techniques. These decisions were made to aid variable selection and to find the optimum solution for creating clusters. Conversely, it was decided to only use the k-means algorithm for all clustering with the caveat that optimisation of the algorithm was essential. It was argued that the processes involved in preparing a dataset for either variable selection or final clustering were more important, and had a greater impact on the final clusters, than which clustering method was to be used. While different clustering algorithms would produce some variation, current and past users identified the top-down three-tiered hierarchical structure of the 2001 OAC as a useful feature. The selection of appropriate variables and the subsequent application of the optimum preparation techniques on them allow the 2011 OAC to be a better classification, but still resemble the 2001 OAC.

The different principles of the 2001 OAC and the 2011 OAC have meant a shift in focus away from solely creating a geodemographic classification. The main goal of the 2001 OAC was to show that creating a free open geodemographic classification was possible at the highest level of granularity. The 2011 OAC expands on this utilising a methodology that does not just cater for one bespoke dataset. There is an opportunity for the methodology used in the 2011 OAC to be a standard bearer for future geodemographic classifications that use Open Data. As such, care has been taken to ensure the methodology can be used with any dataset to create an optimised, and repeatable, geodemographic classification.

With the guiding principles of the 2011 OAC methodology identified and explained in full, the next stage is to explore the outputs of these processes.

Chapter 7

Creating the 2011 Area Classification for Output Areas

7.1. Introduction

The purpose of this chapter is to outline how the methodology discussed in Chapter 6 was implemented in the creation of the 2011 Area Classification for Output Areas (2011 OAC). The different procedures used to accomplish this task are explained and an overview of the main outputs created is provided.

Section 7.2 details the inputs that were used to create the 2011 OAC, and how the process involved firstly identifying an initial selection of variables, then secondly utilising methods to reduce this number before thirdly, a final variable selection was formalised. Section 7.3 provides an overview of how different rate calculation, transformation and standardisation techniques were applied to the final variable selection in order to identify an optimum dataset to use as the basis of the 2011 OAC. The methods used to finalise the number of clusters to form each of the three tiers of the 2011 OAC hierarchy are also explored.

Section 7.4 explains the two core output categories of the 2011 OAC: descriptive and visual. The descriptive outputs of cluster names and descriptions are discussed, and the benefits of different visual outputs are explored. In addition to the core outputs, some of the niche components likely to be used by advanced users of the classification are also detailed, including the R code used to construct the classification. Finally Section 7.5 summarises how the methodology detailed in Chapter 6 was applied to create the 2011 OAC, and how from a technical point of view, the creation of the classification was a success.

7.2. Inputs

The variables used in the construction of the 2011 Area Classification for Output Areas (2011 OAC) were selected solely from the Output Area (OA) and Small Area (SA) level outputs from the 2011 UK Census. This section details the implementation of the variable selection methods utilised for the 2011 OAC as discussed in Sections 6.4 to 6.6. This was a three-stage process which included; the initial selection of the variables; the reduction of this initial selection to a more manageable number of variables; and finally the further reduction to a final list of variables for clustering.

7.2.1. Initial Variable Selection

The initial reduction of the thousands of potential variables available from the 2011 UK Census Key and Quick Statistics released for England and Wales, Scotland and Northern Ireland was performed manually using the variable selection for the 2001 OAC as a guide (see Section 6.4.1). The 2001 OAC initially considered 94 variables for inclusion. These were selected after Vickers (2006) performed a comparative analysis of the types of variables used in previous geodemographic classifications, which were divided into five core domains: Demographic, Household Composition, Housing, Socio-Economic and Employment. By keeping the same five domains for the 2011 OAC a similar selection of variables could be made for the 2011 OAC. In total, 167 initial variables were selected for the 2011 OAC. Table 7.1 details these 167 variables with each colour representing a different domain: Demographic (Red), Household Composition (Blue), Housing (Green), Socio-Economic (Purple) and Employment (Orange).

Table 7.1: The 167 variables initially considered for the 2011 OAC

Code	Variable Name	ST
u001	Males	2
u002	Females	1
u003	Persons living in a household	3
u004	Persons living in a communal establishment	2
u005	Area size (in hectares)	2
u006	Number of persons per hectare	4
u007	Persons aged 0 to 4	3
u008	Persons aged 5 to 9	3
u009	Persons aged 10 to 14	3

Code	Variable Name	ST
u010	Persons aged 15 to 19	1
u011	Persons aged 20 to 24	1
u012	Persons aged 25 to 29	1
u013	Persons aged 30 to 44	1
u014	Persons aged 45 to 59	3
u015	Persons aged 60 to 64	3
u016	Persons aged 65 to 74	3
u017	Persons aged 75 to 84	1
u018	Persons aged 85 to 89	4
u019	Persons aged 90 and over	4
u020	Mean age	1
u021	Median age	1
u022	Persons aged over 16 who are single	1
u023	Persons aged over 16 who are married	3
u024	Persons aged over 16 who are in a registered same-sex civil partnership	2
u025	Persons aged over 16 who are separated	4
u026	Persons aged over 16 who are divorced or formerly in a same-sex civil partnership which is now legally dissolved	3
u027	Persons aged over 16 who are widowed or a surviving partner from a same-sex civil partnership	1
u028	Persons who are white British and Irish	4
u029	Persons who are other white	4
u030	Persons who have mixed ethnicity or are from multiple ethnic groups	4
u031	Persons who are Asian/Asian British: Indian	2
u032	Persons who are Asian/Asian British: Pakistani	2
u033	Persons who are Asian/Asian British: Bangladeshi	2
u034	Persons who are Asian/Asian British: Chinese	2
u035	Persons who are Asian/Asian British: Other	2
u036	Persons who are Black/African/Caribbean/Black British	2
u037	Persons who are Arab or are from another ethnic group	2
u038	Persons who are Christian	3
u039	Persons who are from another religion	4
u040	Persons who have no religion	3
u041	Persons who did not state their religion	3
u042	Persons whose country of birth is the United Kingdom	1
u043	Persons whose country of birth is Ireland	4
u044	Persons whose country of birth is in the old EU (pre 2004 accession countries)	4

Code	Variable Name	ST
u045	Persons whose country of birth is in the new EU (post 2004 accession countries)	2
u046	Persons whose country of birth is not the UK, Ireland or EU countries	4
u047	Persons whose main language is English or their main language is not English but can speak English very well	1
u048	Persons whose main language is not English but can speak English well	4
u049	Persons whose main language is not English and cannot speak English well	4
u050	Persons whose main language is not English and cannot speak English	2
u051	Households that only contain Persons aged over 16 who are living in a couple: Married	3
u052	Households that only contain Persons aged over 16 who are living in a couple: Cohabiting (opposite-sex)	3
u053	Households that only contain Persons aged over 16 who are living in a couple: In a registered same-sex civil partnership or cohabiting (same-sex)	4
u054	Households that only contain Persons aged over 16 who are not living in a couple: Single (never married or never registered a same-sex civil partnership)	1
u055	Households that only contain Persons aged over 16 who are not living in a couple: Married or in a registered same-sex civil partnership	4
u056	Households that only contain Persons aged over 16 who are not living in a couple: Separated (but still legally married or still legally in a same-sex civil partnership)	4
u057	Households that only contain Persons aged over 16 who are not living in a couple: Divorced or formerly in a same-sex civil partnership which is now legally dissolved	1
u058	Households that only contain Persons aged over 16 who are not living in a couple: Widowed or surviving partner from a same-sex civil partnership	1
u059	One person households: Aged 65 and over	3
u060	One person households: Other	1
u061	One family households: All aged 65 and over	4
u062	One family households: Married or same-sex civil partnership couple with no children	3
u063	One family households: Married or same-sex civil partnership couple with dependant children	3
u064	One family households: Married or same-sex civil partnership couple with non-dependant children	4
u065	One family households: Cohabiting couple with no children	4
u066	One family households: Cohabiting couple with dependant children	4
u067	One family households: Cohabiting couple with non-dependant children	2
u068	One family households: Lone parent with dependant children	4
u069	One family households: Lone parent with non-dependant children	4
u070	Other household types: With dependant children	4
u071	Other household types: All full-time students	2
u072	Other household types: All aged 65 and over	2
u073	Other household types: Other	4
u074	Households with no adults in employment: With dependant children	4
u075	Households with no adults in employment: No dependant children	3
u076	Households with lone parent in part-time employment	4
u077	Households with lone parent in full-time employment	4
u078	Households with lone parent not in employment	4

Code	Variable Name	ST
u079	One person ethnic household	3
u080	Household members all have the same ethnic group	3
u081	Households with different ethnic groups between the generations only	4
u082	Households with different ethnic groups within partnerships (whether or not different ethnic groups between generations)	4
u083	Households with any other combination of multiple ethnic groups	4
u084	Household spaces with at least one usual resident	2
u085	Household spaces with no usual residents	4
u086	Households who live in a detached house or bungalow	4
u087	Households who live in a semi-detached house or bungalow	3
u088	Households who live in a terrace or end-terrace house	4
u089	Households who live in a flat	4
u090	Households who live in a caravan or other mobile or temporary structure	2
u091	Households who own or have shared ownership of property	3
u092	Households who are social renting	4
u093	Households who are private renting	1
u094	Households who are living rent free	4
u095	Households who have two or more rooms than required	3
u096	Households who have one more room than required	3
u097	Households who have the required number of rooms	1
u098	Households who have one fewer room than required	4
u099	Households who have two fewer or less rooms than required	2
u100	Households with up to 0.5 persons per room	1
u101	Households with over 0.5 and up to 1.0 persons per room	3
u102	Households with over 1.0 and up to 1.5 persons per room	2
u103	Households with over 1.5 persons per room	2
u104	Day-to-day activities limited a lot or a little Standardised Illness Ratio	1
u105	Persons in very good health	2
u106	Persons in good health	1
u107	Persons in fair health	1
u108	Persons in bad health	1
u109	Persons in very bad health	4
u110	Persons providing unpaid care	3
u111	Persons aged over 16 who have no qualifications	1
u112	Persons aged over 16 whose highest level of qualification is Level 1, Level 2 or Apprenticeship	1
u113	Persons aged over 16 whose highest level of qualification is Level 3 qualifications	1

Code	Variable Name	ST
u114	Persons aged over 16 whose highest level of qualification is Level 4 qualifications and above	1
u115	Persons aged over 16 who are schoolchildren or full-time students	1
u116	Households with no cars or vans	1
u117	Households with 1 car or van	3
u118	Households with 2 or more cars or vans	3
u119	Persons aged between 16 and 74 who work mainly at or from home	1
u120	Persons aged between 16 and 74 who use public transport to get to work	1
u121	Persons aged between 16 and 74 who use private transport to get to work	3
u122	Persons aged between 16 and 74 who walk, cycle or use an alternative method to get to work	1
u123	Persons aged between 16 and 74 who are economically active: Part-time employees	3
u124	Persons aged between 16 and 74 who are economically active: Full-time employees	3
u125	Persons aged between 16 and 74 who are economically active: Self-employed	3
u126	Persons aged between 16 and 74 who are economically active: Unemployed	1
u127	Persons aged between 16 and 74 who are economically active: Full-time student	4
u128	Persons aged between 16 and 74 who are economically inactive: Retired	3
u129	Persons aged between 16 and 74 who are economically inactive: Student (including full-time students)	1
u130	Persons aged between 16 and 74 who are economically inactive: Looking after home or family	1
u131	Persons aged between 16 and 74 who are economically inactive: Long-term sick or disabled	4
u132	Persons aged between 16 and 74 who are economically inactive: Other	4
u133	Persons aged between 16 and 24 who are unemployed	4
u134	Persons aged between 50 and 74 who are unemployed	4
u135	Persons aged between 16 and 74 who have never worked	4
u136	Persons aged between 16 and 74 who are long-term unemployed	4
u137	Employed persons aged between 16 and 74: Part-time working 15 hours or less	3
u138	Employed persons aged between 16 and 74: Part-time working 16 to 30 hours	3
u139	Employed persons aged between 16 and 74: Full-time working 31 to 48 hours	3
u140	Employed persons aged between 16 and 74: Full-time working 49 or more hours	1
u141	Employed persons aged between 16 and 74 industry: Agriculture, forestry and fishing	2
u142	Employed persons aged between 16 and 74 industry: Mining and quarrying	2
u143	Employed persons aged between 16 and 74 industry: Manufacturing	1
u144	Employed persons aged between 16 and 74 industry: Electricity, gas, steam and air conditioning supply	2
u145	Employed persons aged between 16 and 74 industry: Water supply; sewerage, waste management and remediation activities	4
u146	Employed persons aged between 16 and 74 industry: Construction	3
u147	Employed persons aged between 16 and 74 industry: Wholesale and retail trade; repair of motor vehicles and motor cycles	3

Code	Variable Name	ST
u148	Employed persons aged between 16 and 74 industry: Transport and storage	1
u149	Employed persons aged between 16 and 74 industry: Accommodation and food service activities	1
u150	Employed persons aged between 16 and 74 industry: Information and communication	4
u151	Employed persons aged between 16 and 74 industry: Financial and insurance activities	4
u152	Employed persons aged between 16 and 74 industry: Real estate activities	4
u153	Employed persons aged between 16 and 74 industry: Professional, scientific and technical activities	4
u154	Employed persons aged between 16 and 74 industry: Administrative and support service activities	3
u155	Employed persons aged between 16 and 74 industry: Public administration and defence; compulsory social security	1
u156	Employed persons aged between 16 and 74 industry: Education	1
u157	Employed persons aged between 16 and 74 industry: Human health and social work activities	3
u158	Employed persons aged between 16 and 74 industry: Other industry	1
u159	Employed persons aged between 16 and 74 occupation: Managers, directors and senior officials	3
u160	Employed persons aged between 16 and 74 occupation: Professional occupations	3
u161	Employed persons aged between 16 and 74 occupation: Associate professional and technical occupations	3
u162	Employed persons aged between 16 and 74 occupation: Administrative and secretarial occupations	3
u163	Employed persons aged between 16 and 74 occupation: Skilled trades occupations	3
u164	Employed persons aged between 16 and 74 occupation: Caring, leisure and other service occupations	3
u165	Employed persons aged between 16 and 74 occupation: Sales and customer service occupations	3
u166	Employed persons aged between 16 and 74 occupation: Process, plant and machine operatives	3
u167	Employed persons aged between 16 and 74 occupation: Elementary occupations	3

Consideration of a higher number of initial variables for the 2011 OAC in comparison to the 2001 OAC allowed for the inclusion of new outputs from the 2011 UK Census and for the assessment of variables discarded by other geodemographic classifications. In addition, this more extensive selection of initial variables facilitated in addressing concerns raised in the 2011 OAC user engagement (see Section 4.3.2) that the distinction between certain areas of the UK was poor in the 2001 OAC. Testing a greater number of initial variables allowed for greater potential to select those which offered the best differentiation between areas across the UK.

The 167 variables selected were all initially used in their raw form with the exception of u104 ('Day-to-day activities limited a lot or a little Standardised Illness Ratio'). At smaller level geographies, such as OAs or SAs, the use of raw counts or percentages of those who suffer from some form of limiting long term illness is considered unsatisfactory (Vickers

et al., 2005). The age structure of an area can have a significant impact on illness rates. Areas which contain a high concentration of older individuals are more susceptible to having higher illness rates than an area containing a high proportion of younger people. The larger the spatial unit, the more difficult it becomes to distinguish between population attributes, therefore the smaller nature of OAs and SAs make it easier to identify areas dominated by older individuals. It is therefore necessary to standardise illness rates to negate the impact of age structure in each area.

The indirect Standardised Illness Ratio (SIR) method was used by Vickers (2006) in the creation of the 2001 OAC, and works by comparing the observed illness count of an area with the expected value. This comparison is usually performed for multiple age bands, such as '0 to 4', '5 to 9' and so on. The expected count is derived from illness rates of the different age bands across the country. These respective values are then combined across all age brackets so both the total number of those actually ill and those expected to be ill is known for each area. This allows the SIR to be calculated using:

$$SIR^{i} = 100 * (I^{i}/\Sigma_{a}\, r_{a}^{n}\, P_{a}^{i}) \quad (7.1)$$

Where I^{i} equals observed count of ill people in area i, r_{a}^{n} equals rate of illness for age group a in the national population and P_{a}^{i} equals population in area i of age group a. The SIR is relative measure, with an illness rate of 100 being the only constant across the UK. A value of 75 equates to an OA experiencing 25% less illness compared to the national average, and a value of 125 equating to 25% more. Across the UK values ranged from 448.2 to 0. Areas with SIRs below 70 were considered to be the healthiest and those with SIRs above 130, the least healthy (Vickers et al., 2005).

The data available from the 2011 UK Census on the day-to-day activities of the population being limited a lot or a little was provided only as a total for each OA, or the total for those aged between 16 and 64. It was therefore not possible to perform any comparison between multiple age bands making it necessary to adapt the SIR method. Two different age bands were used in the construction of the SIR for the 2011 OAC: 'Aged 0 to 15 and 65 and over'; and 'Aged 16 to 64'. Although this offered a more crude approximation of the healthiness for each OA, the implications of using this measure over the use of a more comprehensive dataset were limited. The age bands used allowed OAs with higher ageing populations to be identified and accounted for.

7.2.2. Reducing Initial Variable Selection

Prior to reduction of the initial variable selection, it was necessary to prepare the variables via rate calculation, transformation and standardisation techniques (see Section 6.5). This created multiple datasets, which were each individually tested to ascertain which variables to remove. The variables were then reduced using four procedures: Pearson's r correlation analysis, within-cluster sum of squares (WCSS) analysis, skewness and the geographic distribution of the variables (see Section 6.6 for detailed explanations of these techniques).

The quantitative nature of the correlation, WCSS and skewness methods provided empirical evidence as to which variables should be kept, merged or removed. The correlation analysis classed values above +0.6 and less than -0.6 as being significant. Whilst no clear thresholds can be used for WCSS analysis, variables can be clearly identified using the technique that have an adverse impact on homogeneity of the final clusters. Performing WCSS analysis on each of the datasets created allowed for variables that repeatedly had a negative impact on cluster homogeneity to be identified.

In contrast, the geographic distribution of the variables in 25 urban locations across the UK provided a more qualitative perspective. Although empirically derived, the relative merits of the geographic distribution of each variable are open to interpretation.

7.2.2.1. Correlation analysis

As discussed in Section 6.6.1, highly correlated variables are not desirable in any geodemographic classification due to the creation of redundancy within the dataset (Ojo et al., 2012). Figure 7.1 is an example correlation matrix of all the 167 variables displayed in Table 7.1 (ordered from 1 to 167), showing where this redundancy existed within the dataset. The same analysis was performed on each of the 27 datasets, and common highly correlated variables identified. Figure 7.1 indicates that the level of correlation that exists between variables in the dataset differs significantly. Instances of correlation range from high inter-variable correlation to variables demonstrating almost no relationship with any other in the dataset. For the purposes of data reduction, the variables of importance are those which indicate significant correlation (values above +0.6 and less than -0.6). These instances are shown in Figure 7.2, an example correlation matrix for one of the 27 datasets of variables which only exhibit significant correlation.

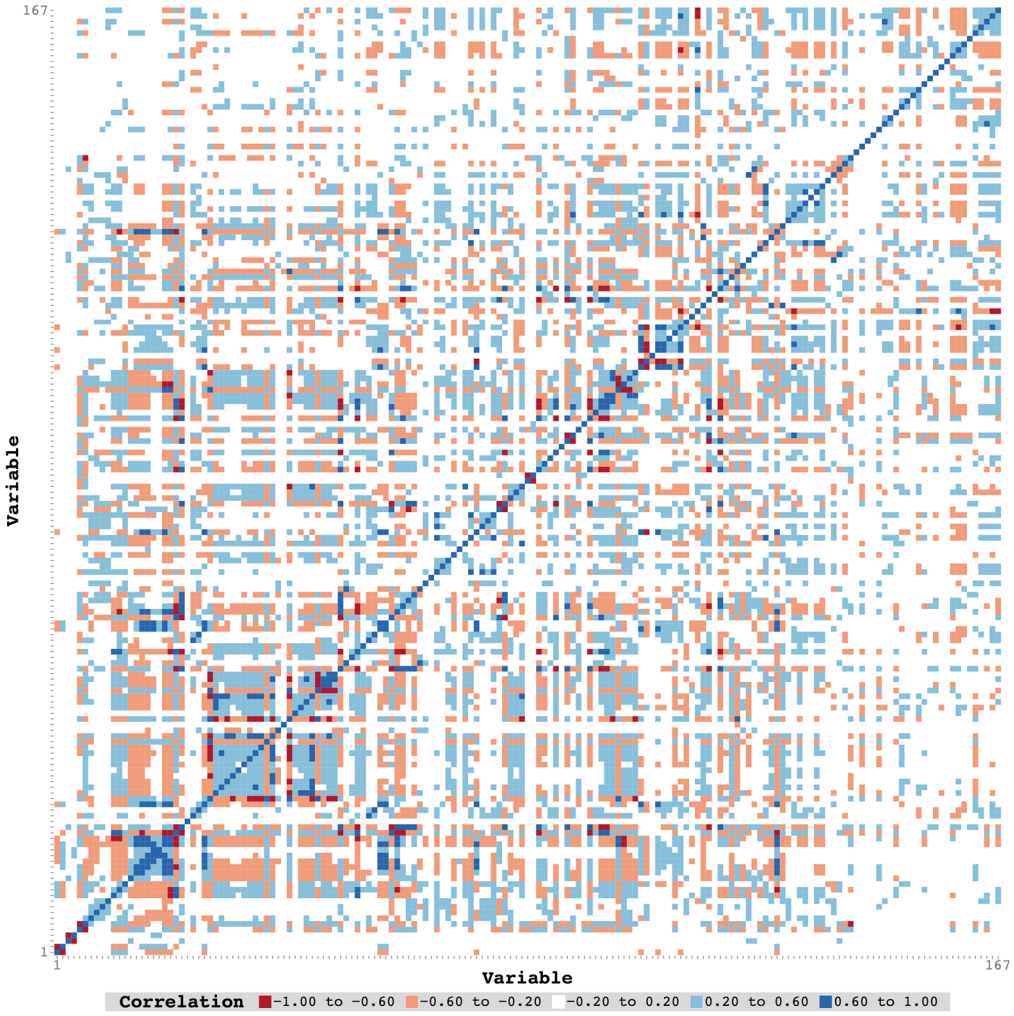


Figure 7.1: Correlation matrix of the 2011 OAC's 167 initially selected variables (Ordered from variable u001 to u167 - see Table 7.1 for variable names)

Figure 7.2 reveals a sporadic distribution of positively and negatively inter-correlated variables. Certain variables are correlated with several others, while in other instances a variable will only correlate with one or two in the dataset. The utilisation of correlation analysis to reduce the size of the dataset requires an understanding of the different types of relationships that exist between the variables. The patterns of correlation seen in Figure 7.2 were also found while performing similar analysis in the reduction of the initial variable selection for the 2001 OAC. Vickers (2006) noted that three different types of correlation existed. The first are variables which share the same denominator, therefore an individual can only be allocated to one variable. This results in strong negative correlations between these types of variables, which will be perfect (-1) if only two are present. In these instances, variables can be removed whilst maintaining the ability to differentiate. For example by retaining the 'Male' variable and removing

'Female', you do not lose the ability to identify areas where there are high concentrations of females as these will be in locations where the male concentrations are low.

The second type of correlation identified by Vickers (2006) are where variables are inherently connected due to causality. These are instances where variables are based on a common theme, but do not have the same denominator. This is prevalent in the 2011 UK Census. For example, it is possible for variables on housing to be released at the individual or household level. The use of a different denominator does not stop them being related, and while the relationship between these types of variables will vary on a case-by-case basis, it is likely that a strong link will exist.

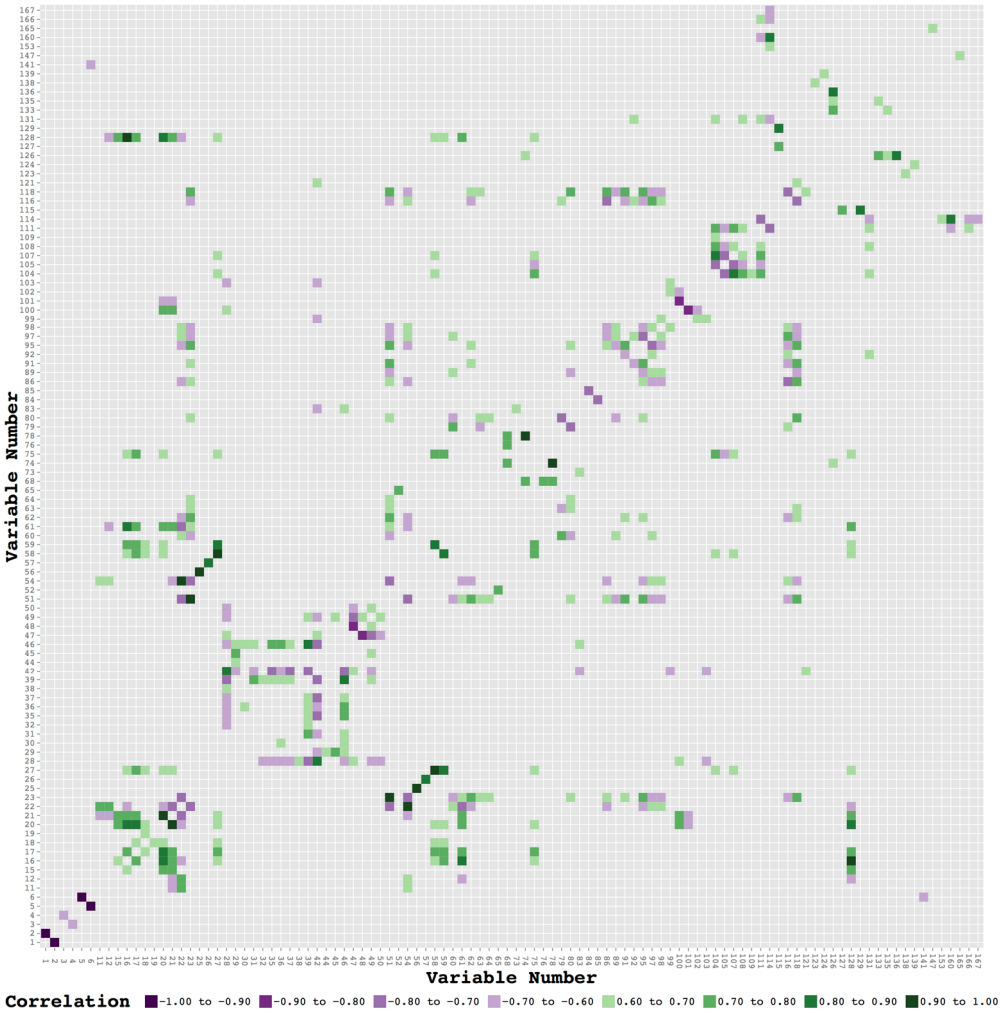


Figure 7.2: Significant correlation matrix of the 2011 OAC's 167 initially selected variables

(See Table 7.1 for variable names)

The third type of correlation is found between variables "where the presence of one [variable] indicates the presence or absence of another, but does not fundamentally cause it to be so" (Vickers et al., 2005, p. 9). An example of this type of relationship, seen in Figure 7.2, is between persons born outside of the European Union and persons who are not Christian, where a strong positive correlation between these two variables exists. As such, it makes it likely that an individual would respond either 'yes' or 'no' to both questions, despite this not having to be the case.

Vickers et al. (2005) described this third type of correlation as the most interesting, as these relationships between variables are not preordained and the inclusion of such variables can enhance the predictive and descriptive power of a classification. Despite this potential, the removal of highly correlated variables from the 2011 OAC, either by removing them completely or merging them to create composite variables, was key in the creation of a final dataset. This reduced data redundancy and avoided giving certain variables too much prominence within the final classification.

7.2.2.2. Within-cluster sum of squares analysis

The within-cluster sum of square analysis (WCSS) technique discussed in Section 6.6.3 was used to identify badly performing variables from the 167 initially selected. Figure 7.3 is an example output of WCSS analysis using the 'Percentages, Inverse Hyperbolic Sine, Range' dataset. Several variables can be identified as having adverse impacts on the cluster analysis, but 'households who live in a flat' can be highlighted as having a notably negative impact on the cluster analysis. The same analysis was performed on each of the 27 unique 'rate calculated, transformed and standardised' datasets. At this stage in the creation of the 2011 OAC it was not known which combination of these procedures would be used on the dataset for clustering. Therefore, it would not have been appropriate to base variable selection on WCSS analysis performed on a single dataset alone.

The WCSS analysis performed on each of the 27 datasets revealed the worst performing variables. The ten variables from each dataset that had the greatest negative impact on cluster homogeneity were identified. This revealed that variables from certain domains were more likely to have a negative impact on cluster homogeneity than others. The 'Housing' domain for example had on average 8.8% of its variables from each dataset identified as being in the top ten worst performers, compared to 6.5% in the 'Housing

Composition' domain, 6.2% in the 'Demographic' domain, 5.3% in the 'Employment' domain and 2.8% from the 'Socio-Economic' domain.

The greater propensity for variables from the 'Housing' and 'Housing Composition' domains to perform badly is also seen when looking at which variables repeatedly had the greatest negative impact on cluster homogeneity. The total number of times a variable was identified as being in the top ten worst performers for each dataset was calculated, with Table 7.2 highlighting the eleven individual variables considered to be the worst performing. The variable 'persons living in a communal establishment' was identified as the worst potential variable for the 2011 OAC, as it was in the top ten worst performing variables in 10 of the 27 datasets. The domains represented by the eleven worst performing variables were however mixed. An indication that individual variables, no matter which domain they are assigned to, could still have a negative impact on overall cluster homogeneity.

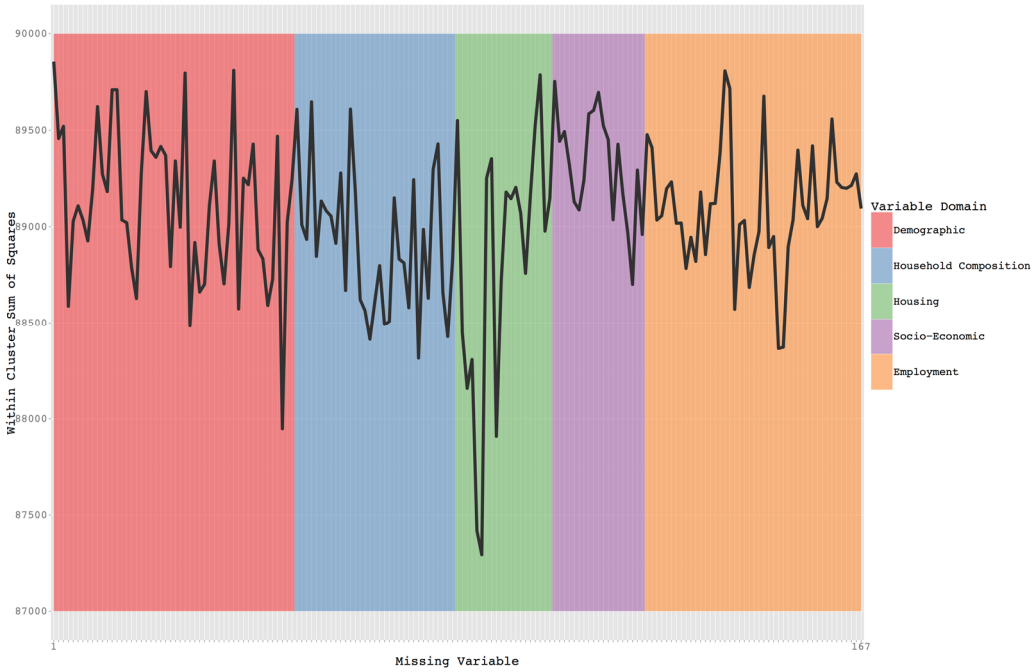


Figure 7.3: Missing variables WCSS values for the 2011 OAC's 167 initially selected variables

(Ordered from variable u001 to u167 - see Table 7.1 for variable names)

Table 7.2: The eleven worst performing variables using the WCSS analysis technique across the 27 uniquely converted, transformed and standardised datasets

Code	Variable Name	Variable Domain	Occurrences
u004	Persons living in a communal establishment	Demographic	10
u067	One family households: Cohabiting couple with non-dependant children	Household Composition	8
u090	Households who live in a caravan or other mobile or temporary structure	Housing	8
u144	Employed persons aged between 16 and 74 industry: Electricity, gas, steam and air conditioning supply	Employment	8
u072	Other household types: All aged 65 and over	Household Composition	7
u094	Households who are living rent free	Housing	6
u003	Persons living in a household	Demographic	5
u024	Persons aged over 16 who are in a registered same-sex civil partnership	Demographic	5
u033	Persons who are Asian/Asian British: Bangladeshi	Demographic	5
u041	Persons who did not state their religion	Demographic	5
u145	Employed persons aged between 16 and 74 industry: Water supply; sewerage, waste management and remediation activities	Employment	5

The reason for the bad performance of these variables is likely to be a result of the geographic variation in the interactions they have with other variables. The 'persons living in a communal establishment' can be explained by the variations in the areas where communal establishments are found. They can be found in both areas of deprivation and affluence across the UK. The characteristics of the people that live in this type of housing are therefore not consistent, which has an adverse impact on the clustering process. As a clustering algorithm looks for patterns in the data, any variable that has a diverse relationship with other variables across geographic space creates more heterogeneous clusters in order to accommodate this variation.

Interpreting the WCSS analysis to aid the reduction of variables did not automatically mean removing those that performed badly. As shown in Sections 6.8.6 and 6.8.7, the WCSS value can be used to aid the formation of an optimum clustering outcome, albeit influenced by the geographical extent of the data. As such, the variables identified in Figure 7.3 and Table 7.2 as performing badly reflect only the entire UK dataset. Certain variables are universal in how they describe an area (such as those relating to age), while others less so. A high concentration of a particular housing type in an affluent area for example means something different to similar concentrations in more deprived areas. These types of variations are required for a geodemographic classification to create unique clusters. The WCSS analysis does however, suggest that a variable can have too much variation, to the extent where cluster homogeneity is impacted. A decision therefore needs to be made between retention of the differentiation power offered by a variable and producing the most homogeneous clusters possible. Either option can be justified within the wider context of creating an optimum 2011 OAC, although it means that optimum cannot refer to both cluster composition and variable differentiation. As such, each variable not considered suitable from a WCSS perspective was either excluded or retained from the 2011 OAC final selection on a variable-by-variable basis.

7.2.2.3. Skewed variables

The skewness of individual variables differs greatly. An example of this is shown with the 167 variables from the 'Percentages, Inverse Hyperbolic Sine, Range' dataset shown in Figure C.1 in Appendix C. The distributions of individual variables range from highly negatively skewed to highly positively skewed; certain variables have a more normal distribution, and others are bimodal. Although the variation is great, some patterns can be seen between variables that belong to the different domains. A large number of variables in the 'Demographic' domain are highly skewed, both positively and negatively. This would suggest that certain variables are likely to be dispersed across the UK, such as 'persons who live in a household', while others are more likely to be concentrated in smaller geographical areas, such as 'persons who are Asian/Asian British: Bangladeshi'. Some variables, such as 'persons who have mixed ethnicity or are from multiple ethnic groups' have bimodal distributions, suggesting individuals who have mixed ethnicity or are from multiple ethnic groups are present in both high and low concentrations across different areas of the UK.

Variables in the 'Socio-Economic' domain appear to be more normally distributed, suggesting that the majority of variables in this domain are commonplace across the UK and distributed fairly evenly across all areas. An exception to this is the 'households with one car or van' variable, which has a strong negative distribution. The distribution of variables in the 'Housing' domain are more mixed, with some, such as 'households who are private renting', having an almost normal distribution, while others like 'households who own or have shared ownership of property' having a negative distribution. This would suggest that certain housing indicators are distributed fairly evenly across all of the UK, while others are more likely to be found in smaller geographical areas. The majority of variables in the 'Household Composition' domain appear to contain bimodal distributions, suggesting the prevalence of certain variables to have both high and low concentrations across different areas of the UK. Finally, variables in the 'Employment' domain either have a slight negative or positive skew, suggesting that most industries do not have an excessive high or low number of employees per OA or SA across the UK.

Based on the skewed nature of the distribution of the 167 variables across the 27 datasets, each variable can be categorised into one of four groups:

1. **Normal distribution**. Variables are likely to be commonplace across the UK and be distributed fairly evenly across all areas. Variables described by this category will either have a perfect normal distribution, or a marginal positive or negative skew. **42 in total**.
2. **Positive distribution**. Variables are more likely to be found in higher concentrations in a limited number of areas across the UK. Other areas in the UK that contain these variables are likely to be fewer in number and contain smaller concentrations. **25 in total**.
3. **Negative distribution**. Variables are more likely to be found in lower concentrations in a large number of areas across the UK. Other areas in the UK that contain these variables are likely to be fewer in number, but contain larger concentrations. **49 in total**.
4. **Bimodal distribution**. Variables that are found both in either high or low concentrations across a limited number of areas in the UK and also have a distribution that closer reflects normality (and therefore considered more commonplace) across other parts of the country. **51 in total**.

The column 'ST' (Skewness Type) in Table 7.1 details the group to which each variable belongs. As discussed in Section 6.6.4, highly skewed variables were kept in the dataset.

These variables show the absence or dominance of a variable in an area and have the potential to indicate significant traits, allowing for differentiation between areas.

7.2.2.4. Geographic distribution of population characteristics

As discussed in Section 6.6.5, the inclusion of variables that did not have uniform geographic distribution was desirable for the 2011 OAC. The distribution of the 167 variables initially selected for inclusion in the 2011 OAC were analysed across 25 towns and cities in the UK. The list of the towns and cities used is shown in Section C.2 in Appendix C. An example of this analysis showed the 'households who live in a terrace or end-terrace house' variable ranged from representing 49% of the population in Kingston upon Hull to 12% in Glasgow. A non-uniform geographic distribution was also prevalent across other variables; 23 variables had a range of over 20% between their highest and lowest concentrations in the 25 locations analysed, while 49 variables had a range of over 10%.

The largest percentage of the population represented by the variable, 'persons who are Asian/Asian British: Bangladeshi', was 3% in Birmingham, with the lowest being in Belfast and Plymouth at 0.1%, a range of 2.9%. This range is comparatively small when comparing it with the Indian and Pakistani equivalents of the variable, 28% and 20% respectively. The removal of variables that have such a small range, and therefore limited differentiation across the UK, would appear to be beneficial for the 2011 OAC. Retention of the Indian and Pakistani variables and exclusion of the Bangladeshi variable would, however be an over simplification of the respective merits of the individual variables. In certain cases, the limited differentiation offered is a result of the variable representing only a small percentage of the overall population of the UK.

Interpreting a variable based on its concentration in urban areas is inevitably influenced by the chosen geography. A city such as Birmingham consists of 3,223 OAs, and this relatively large area can hide many small area variations. For example, looking at Birmingham as a whole, it is not possible to determine whether the 3% of persons who identify themselves as Bangladeshi are concentrated to a small number of OAs, or are dispersed across the entire city. This can be calculated if alternative geographies were used, such as Wards, but this would make the resulting outputs too complicated to interpret from a variable reduction perspective. Variables with these types of

distribution characteristics therefore required consideration on an individual basis for the 2011 OAC. A variable which represents a small proportion of the population and is found in concentrated locations suggests that it is an important characteristic of that area. Conversely, the importance of a variable diminishes if it is dispersed across a wider geographic area.

7.2.3. Final Variable selection

The techniques used to aid variable reduction and the interpretation of the results of these processes, as discussed in Section 7.2.2, led to the formulation of a final list of variables for the 2011 OAC. From the 167 variables initially selected: 84 were removed; 40 were retained in their original form; and 43 were merged, creating 20 composite variables. The resulting 60 variables were used to create the 2011 OAC. This is 46% more variables than that were used in the 2001 OAC, reflecting the greater range of outputs provided by the 2011 UK Census. A full list of the 60 variables shown in Table 7.3 with a breakdown of the domains and subdomains totals shown in Table 7.4.

Table 7.3: The 60 variables selected to create the 2011 OAC

Code	Variable Name	Domain	Subdomain
k001	Persons aged 0 to 4	Demographic	Population Age
k002	Persons aged 5 to 14	Demographic	Population Age
k003	Persons aged 25 to 44	Demographic	Population Age
k004	Persons aged 45 to 64	Demographic	Population Age
k005	Persons aged 65 to 89	Demographic	Population Age
k006	Persons aged 90 and over	Demographic	Population Age
k007	Number of persons per hectare	Demographic	Population Structure
k008	Persons living in a communal establishment	Demographic	Population Structure
k009	Persons aged over 16 who are single	Demographic	Marital and Civil Partnership Status
k010	Persons aged over 16 who are married or in a registered same-sex civil partnership	Demographic	Marital and Civil Partnership Status
k011	Persons aged over 16 who are divorced or separated	Demographic	Marital and Civil Partnership Status
k012	Persons who are white	Demographic	Ethnicity

Code	Variable Name	Domain	Subdomain
k013	Persons who have mixed ethnicity or are from multiple ethnic groups	Demographic	Ethnicity
k014	Persons who are Asian/Asian British: Indian	Demographic	Ethnicity
k015	Persons who are Asian/Asian British: Pakistani	Demographic	Ethnicity
k016	Persons who are Asian/Asian British: Bangladeshi	Demographic	Ethnicity
k017	Persons who are Asian/Asian British: Chinese and Other	Demographic	Ethnicity
k018	Persons who are Black/African/Caribbean/Black British	Demographic	Ethnicity
k019	Persons who are Arab or from other ethnic groups	Demographic	Ethnicity
k020	Persons whose country of birth is the United Kingdom or Ireland	Demographic	Region of Birth
k021	Persons whose country of birth is in the old EU (pre 2004 accession countries)	Demographic	Region of Birth
k022	Persons whose country of birth is in the new EU (post 2004 accession countries)	Demographic	Region of Birth
k023	Persons whose main language is not English and cannot speak English well or at all	Demographic	Proficiency in English
k024	Households with no children	Household Composition	Household Type
k025	Households with non-dependant children	Household Composition	Household Type
k026	Households with full-time students	Household Composition	Household Type
k027	Households who live in a detached house or bungalow	Housing	Housing Type
k028	Households who live in a semi-detached house or bungalow	Housing	Housing Type
k029	Households who live in a terrace or end-terrace house	Housing	Housing Type
k030	Households who live in a flat	Housing	Housing Type
k031	Households who own or have shared ownership of property	Housing	Housing Ownership
k032	Households who are social renting	Housing	Housing Ownership
k033	Households who are private renting	Housing	Housing Ownership
k034	Households who have one fewer or less rooms than required	Housing	Housing Crowding
k035	Individuals day-to-day activities limited a lot or a little (Standardised Illness Ratio)	Socio-Economic	Population Health and Care
k036	Persons providing unpaid care	Socio-Economic	Population Health and Care
k037	Persons aged over 16 whose highest level of qualification is Level 1, Level 2 or Apprenticeship	Socio-Economic	Population Education

Code	Variable Name	Domain	Subdomain
k038	Persons aged over 16 whose highest level of qualification is Level 3 qualifications	Socio-Economic	Population Education
k039	Persons aged over 16 whose highest level of qualification is Level 4 qualifications and above	Socio-Economic	Population Education
k040	Persons aged over 16 who are schoolchildren or full-time students	Socio-Economic	Population Education
k041	Households with two or more cars or vans	Socio-Economic	Vehicle Availability
k042	Persons aged between 16 and 74 who use public transport to get to work	Socio-Economic	Travel-to-Work
k043	Persons aged between 16 and 74 who use private transport to get to work	Socio-Economic	Travel-to-Work
k044	Persons aged between 16 and 74 who walk, cycle or use an alternative method to get to work	Socio-Economic	Travel-to-Work
k045	Persons aged between 16 and 74 who are unemployed	Employment	Population Employment
k046	Employed persons aged between 16 and 74 who work part-time	Employment	Employment Hours
k047	Employed persons aged between 16 and 74 who work full-time	Employment	Employment Hours
k048	Employed persons aged between 16 and 74 who work in the agriculture, forestry or fishing industries	Employment	Industry Sector
k049	Employed persons aged between 16 and 74 who work in the mining, quarrying or construction industries	Employment	Industry Sector
k050	Employed persons aged between 16 and 74 who work in the manufacturing industry	Employment	Industry Sector
k051	Employed persons aged between 16 and 74 who work in the energy, water or air conditioning supply industries	Employment	Industry Sector
k052	Employed persons aged between 16 and 74 who work in the wholesale and retail trade; repair of motor vehicles and motor cycles industries	Employment	Industry Sector
k053	Employed persons aged between 16 and 74 who work in the transport or storage industries	Employment	Industry Sector
k054	Employed persons aged between 16 and 74 who work in the accommodation or food service activities industries	Employment	Industry Sector
k055	Employed persons aged between 16 and 74 who work in the information and communication or professional, scientific and technical activities industries	Employment	Industry Sector
k056	Employed persons aged between 16 and 74 who work in the financial, insurance or real estate industries	Employment	Industry Sector
k057	Employed persons aged between 16 and 74 who work in the administrative or support service activities industries	Employment	Industry Sector
k058	Employed persons aged between 16 and 74 who work in the in public administration or defence; compulsory social security industries	Employment	Industry Sector
k059	Employed persons aged between 16 and 74 who work in the education sector	Employment	Industry Sector
k060	Employed persons aged between 16 and 74 who work in the human health and social work activities industries	Employment	Industry Sector

Table 7.4: The number of 2011 OAC variables assigned to the classification domains and subdomains

Domain	Subdomain	Variables
Demographic	Population Age	6
Demographic	Population Structure	2
Demographic	Marital and Civil Partnership Status	3
Demographic	Ethnicity	8
Demographic	Region of Birth	3
Demographic	Proficiency in English	1
Household Composition	Household Type	3
Housing	Housing Type	4
Housing	Housing Ownership	3
Housing	Housing Crowding	1
Socio-Economic	Population Health and Care	2
Socio-Economic	Population Education	4
Socio-Economic	Vehicle Availability	1
Socio-Economic	Travel-to-Work	3
Employment	Population Employment	1
Employment	Employment Hours	2
Employment	Industry Sector	13

The concluding decisions that were made to reduce the 167 initial variables to the final 60 are explained fully in Table C.1 in Appendix C. Although each variable was considered individually, the actions taken to reduce the number of variables are summarised into ten-fold rationale:

1. Variables retained without modification
2. Variables merged to reduce high inter-correlation
3. Variables merged to reduce skewness in the composite variable created
4. Variables merged to improve geographic variation
5. Variables merged for another reason
6. Variables removed due to high inter-correlation
7. Variables removed due to being identified as behaving badly by WCSS analysis
8. Variables removed due to their skewed distribution
9. Variables removed due to poor geographic variation
10. Variables removed for another reason

In certain instances, multiple rationale were applied to the same variable; for example, where a variable was merged with another due to both high inter-correlation and to provide a better geographic representation across the UK. Table 7.5 is a breakdown of how these rationales were applied to the 167 variables (the bullet point numbers correspond to the table numbering). The actions performed on the majority of variables

were done so for a single rationale, although 9 variables were merged and 30 variables removed for multiple rationales. The primary rationale for the removal or merging of variables was inter-correlation. Removing a variable in these circumstances reduced the redundancy within the dataset and had no significant impact on the discriminatory power of the classification. In other cases where variables were merged, such as those relating to age, inter-correlation within the dataset was reduced and allowed for flexibility in the range of the output produced for a particular. This flexibility allowed for retention of distinctions between children (such as pre and post school age), whilst age categories over 65 were split into only two categories; above and below 90 years old. Detailed breakdowns of the population post-retirement age were considered less vital in differentiating different population cohorts across the UK. However, a variable for those aged over 90 was included in the 2011 OAC, as England and Wales can be considered to have an ageing population (ONS, 2012n).

Table 7.5: Reasons for 2011 OAC variable reduction (numbers correspond to numbered bullet points on page 243)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	41									
2	-	9								
3	-	5	1							
4	-	-	2	5						
5	-	2	-	-	16					
6	-	-	-	-	-	31				
7	-	-	-	-	-	-	-			
8	-	-	-	-	-	3	-	-		
9	-	-	-	-	-	6	3	2	8	
10	-	-	-	-	-	13	1	1	-	17

The retention of variables that were either highly skewed, had limited geographic variation or represented only a small proportion of the population were only kept if that particular variable was considered important to the 2011 OAC (as described in Table C.1). In certain cases, highly skewed variables were merged in an attempt to reduce the skewness, which in turn increased the proportion of the population represented. In certain cases it was considered inappropriate to remove variables, even if they were highly correlated, skewed or performed badly in the WCSS analysis, due to their overall importance to the 2011 OAC. Although the desired outcome of the clustering process is to produce the most homogenous clusters possible, this could not be done at the expense of variables key to describing the characteristics of an area. For example, removal of the majority of the housing variables would have resulted in limited indicators of the physical environment of an area.

As a greater importance was placed on the capacity for the retained variables to differentiate between different areas, the 60 variables retained an element of inter-correlation, as shown in Figure 7.4. The amount of significant inter-correlation found in the 60 variables, shown in Figure 7.5, does mean that redundancy remains within the 2011 OAC dataset, albeit at a more acceptable level when compared to that shown in Figure 7.2.

The importance of the capacity for the variables to differentiate between different areas also affected the removal of those with highly skewed distributions. Figure 7.6 shows the mean skewness value for each variable calculated from the 27 datasets created from the different rate calculation, transformation and standardisation methods. This indicates that many of the 60 variables selected retained highly skewed distributions, with only 30 variables demonstrating skewness values between -1 and 1. Variables that had larger skewness values were kept if they had a greater capacity for area differentiation. For example, the variable k058 ('persons who work in the in public administration or defence; compulsory social security industries'), with a skewness value of 8.7, was retained to provide greater variation in the types of industry in the UK. Removal or merging of the variable would have resulted in a more generic representation in the 2011 OAC of the types of industry individuals are employed in. The benefits of retaining variables with a higher inter-correlation and with highly skewed distributions was considered greater than the negative impact of their inclusion on the clustering process.

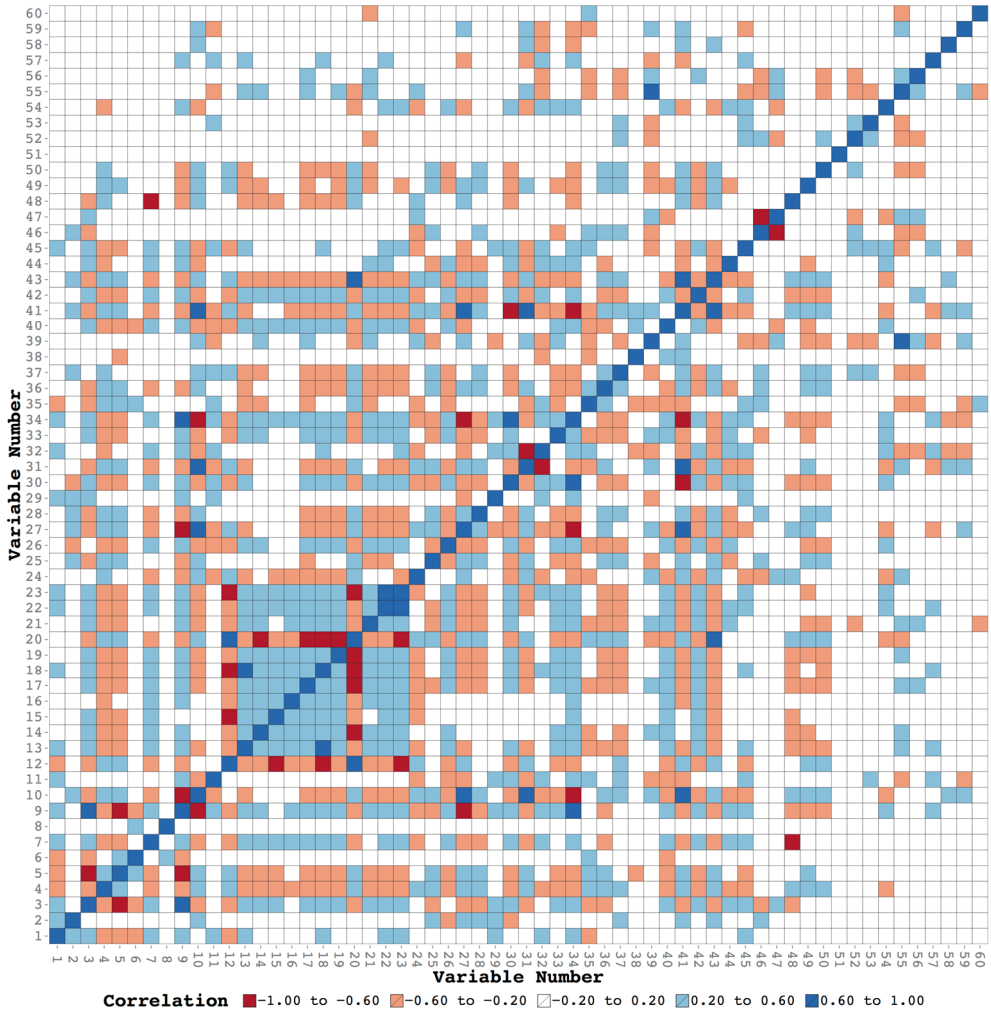


Figure 7.4: Correlation matrix of the 2011 OAC's 60 final selected variables

(See Table 7.3 for variable names)

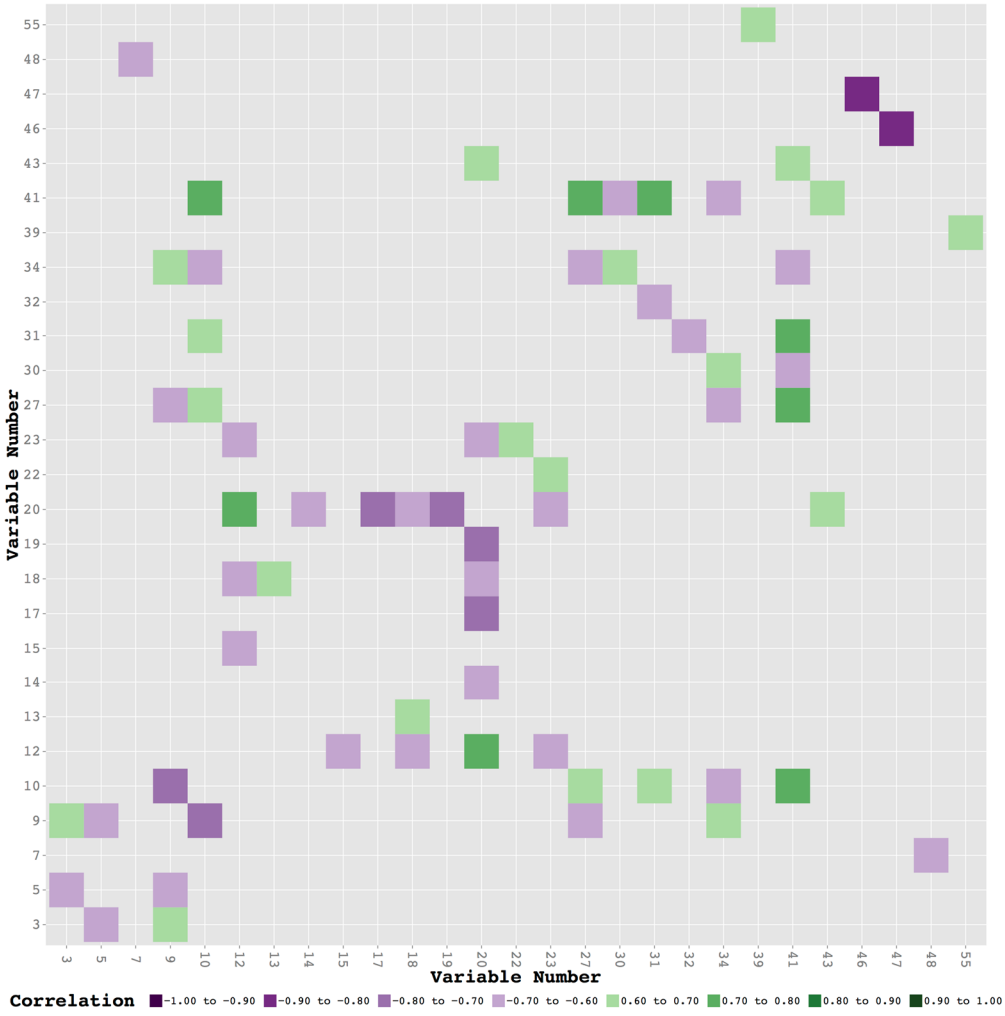


Figure 7.5: Significant correlation matrix of the 2011 OAC's 60 final selected variables

(See Table 7.3 for variable names)

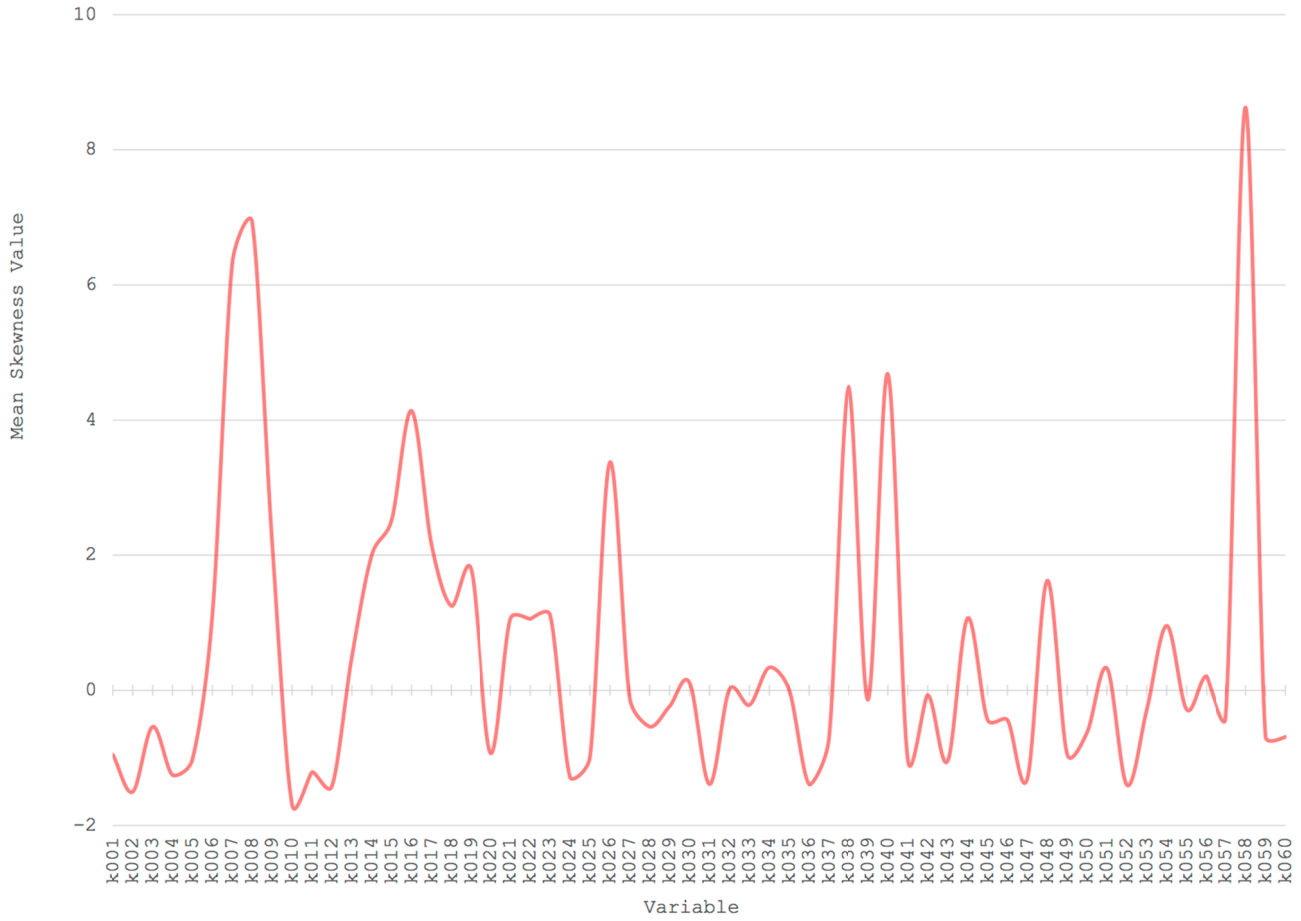


Figure 7.6: The Mean Skewness of the 27 datasets created from the rate calculation, transformation and standardisation techniques (See Table 7.3 for variable names)

Variables with limited geographic variation, however, would have had a greater negative impact with their inclusion in the clustering process than in their retention. Figure 7.7 shows the maximum difference between 25 towns and cities in the UK (see Section C.2) with the lowest and highest concentrations for each of the final 60 variables. As k007 (number of people per hectare) and k035 (day-to-day activities limited a lot or a little standardised illness ratio) are expressed as ratios, they cannot be directly compared with the other 58 variables expressed as percentages. 44 variables have a range of over 5% between their lowest and highest concentrations and 30 have a range of over 10%. Two variables with notable geographic variation are k012 (persons who are white) and k030 (households who live in flats). It is this level of variation in the geographic distribution of variables that allows clustering algorithms to function and identify unique groupings.

For the purposes of selecting the final variables for the 2011 OAC, geographic distribution was considered to be less important than correlation, WCSS and skewness analysis. These techniques provided a greater insight into the non-spatial aspects of each variable and it was considered more important to remove badly performing variables identified using these methods due to the greater negative impact they would have had on the final classification. As such, a variable with good geographic variation was unlikely to be retained if it had high correlation or skewness. Although all of these techniques contributed towards the final selection, if a variable was considered to be important enough to the classification then it was accommodated in the final variable selection.

The final 60 variables selected for the 2011 OAC cannot be considered to be perfect. The variation in the population across the UK is too great for any selection of variables to conform perfectly to any statistical measure. A balance was therefore struck between selecting characteristics which most accurately represent the population of the UK, whilst conforming to the best practices of geodemographic classifications.

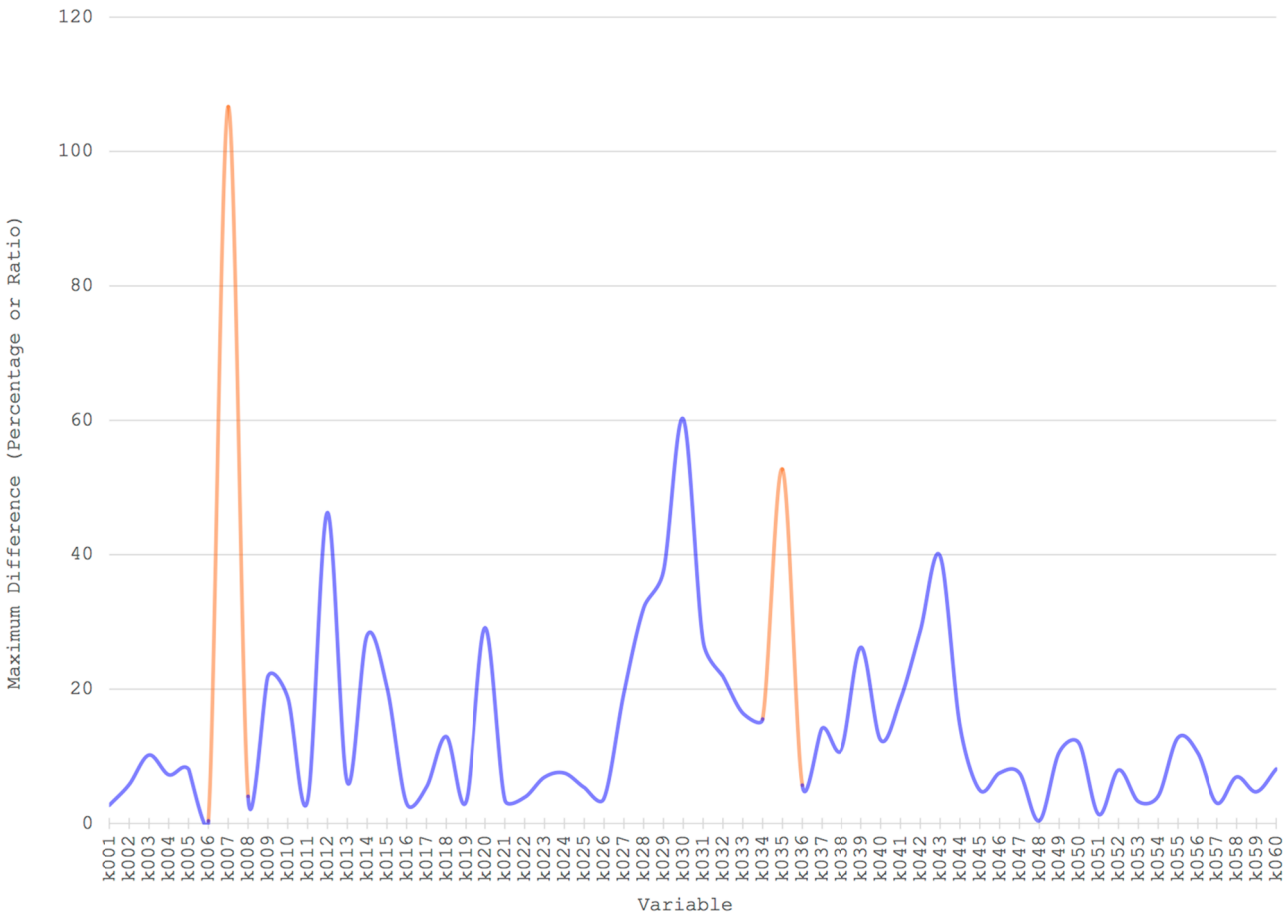


Figure 7.7: Maximum difference in variable distribution between 25 urban areas in the UK

(See Table 7.3 for variable names)

Additionally, the final selection of variables is also a reflection of the subjective choices made by those creating the 2011 OAC. The choices made include: those made during the initial selection of the 167 variables; the selection of the methods used to aid reduction; and decisions regarding which variables should be retained based on their importance to the final classification. These choices will have had a significant influence on the final classification. However, by fully documenting the processes involved, users are able to understand the reasons behind every decision. Due to the subjectivity of decision-making at certain points, it is highly likely that a different creator of the classification would have selected different variables. However, the 'core' components; such as age, ethnicity, employment status along with variables that relate to the physical characteristics of an area (i.e. housing) are likely to have remained.

7.3. Processes

This section details the processes undertaken to create the 2011 OAC from the selected 60 variables. Firstly, an optimum dataset for clustering was selected from the 27 created by the different permutations of the rate calculation, transformation and standardisation techniques described in Section 6.5. Secondly, an optimum number of clusters were identified to form the top level of the classifications hierarchy – the Supergroups.

The selection of the optimum dataset and the number of clusters used in the creation of the Supergroups led to the creation of the lower tiers in the classifications hierarchy. As discussed in Section 6.8.3, the use of the k-means clustering algorithm in the 2011 OAC, produced a three tiered top-down classification, where the Supergroups (top tier) were used to form the Groups (middle tier), and these in turn were used to create the Subgroups (bottom tier).

7.3.1. Identifying optimum dataset and cluster numbers

To achieve the optimum clustering solution for the 2011 OAC, the 'best' dataset and 'best' number of clusters were identified. The optimum dataset was selected based on the performance of each dataset when clustered in terms of the composition, size and geographic distribution of each group. As discussed in Section 6.8.4, the total number of Supergroups was selected from only 5 to 9 cluster solutions.

7.3.1.1. Optimum dataset

Each cluster solution was subsequently applied to each dataset, creating 135 uniquely clustered datasets (five cluster solutions for each of the 27 unique datasets). The suitability of each dataset and cluster solution was then initially assessed in terms of the composition, size and geographic distribution of the groups. Histograms shown in Figure C.2 in Appendix C were used initially to explore the degree to which the datasets were normally distributed. However, three rationales in total were ultimately used to reduce the number of datasets from 27 to 4:

1. Datasets that had skewness values above 1 or less than -1
2. Datasets that led to the creation of 'micro clusters' (where the smallest cluster(s) only accounted for a small percentage of the population)
3. Datasets that led to the creation of clusters that offered little differentiation power across the UK with low levels of cluster homogeneity

Datasets demonstrating at least one of these traits were considered undesirable and were removed from consideration. This permitted performance of a more in-depth analysis on datasets which were considered better suited to creating the 2011 OAC, such as looking at the geographical distribution of the clusters. The results of the dataset reduction are shown in Table 7.6. In total, 12 datasets were removed from consideration because their skewness was above the threshold values. Datasets which created micro clusters totalled seven, although it should be noted that this value was conditional to the subjectivity of when a cluster could be considered as 'micro'. Decisions were made based on the size of the smallest cluster in relation to the others created. A clustering solution which produced multiple smaller clusters was not immediately considered as undesirable. This implies that the groupings found by the k-means clustering algorithm reflect distinct variation in the population. A clustering solution which produced equally sized clusters with one micro cluster was considered as undesirable. The micro cluster was likely to be an artefact of the division of the dataset into n clusters by the k-means algorithm, rather than the identification of distinct groupings within the data. Finally, four datasets were removed because the clusters created offered little differentiation power across the UK. Figure 7.8 provides an example cluster profile created from the 'Mean Difference, Box-Cox, Range' dataset. The red line represents the national average, and the blue line represents how each variable varies in that particular cluster away from this. Where a variable is close to the national average, demonstrates very limited differentiation power of a cluster. Cluster solutions which were predominately formed

of clusters like these were undesirable and were therefore not used in the construction of the 2011 OAC.

Table 7.6: The 27 datasets considered to create the 2011 OAC (Dataset removal reasons correspond to numbered bullet points on page 252)

Dataset Number	Data Modification	Skew	Mean SED value for 6 to 8 Clusters	Dataset Removal Reason
1	Percentages, Box-Cox, Z-Scores	0.22	42.52	2
2	Percentages, Box-Cox, Range	0.39	1.40	Kept
3	Percentages, Box-Cox, Inter-Decile Range	0.31	7.01	2
4	Percentages, Log, Z-Scores	-0.38	50.46	2
5	Percentages, Log, Range	-0.20	0.94	Kept
6	Percentages, Log, Inter-Decile Range	0.13	10.60	3
7	Percentages, Inverse Hyperbolic Sine, Z-Scores	-0.60	48.68	2
8	Percentages, Inverse Hyperbolic Sine, Range	-0.36	0.98	Kept
9	Percentages, Inverse Hyperbolic Sine, Inter-Decile Range	-0.26	10.77	2
10	Index Scores, Box-Cox, Z-Scores	0.63	52.25	2
11	Index Scores, Box-Cox, Range	0.34	0.49	Kept
12	Index Scores, Box-Cox, Inter-Decile Range	0.67	10.11	2
13	Index Scores, Log, Z-Scores	-3.01	79.07	1
14	Index Scores, Log, Range	-1.56	1.63	1
15	Index Scores, Log, Inter-Decile Range	-5.76	63.76	1
16	Index Scores, Inverse Hyperbolic Sine, Z-Scores	-3.17	80.21	1
17	Index Scores, Inverse Hyperbolic Sine, Range	-1.59	1.63	1
18	Index Scores, Inverse Hyperbolic Sine, Inter-Decile Range	-6.02	69.02	1
19	Mean Difference, Box-Cox, Z-Scores	4.27	269.96	1
20	Mean Difference, Box-Cox, Range	0.11	0.05	3
21	Mean Difference, Box-Cox, Inter-Decile Range	80.08	1432.30	1
22	Mean Difference, Log, Z-Scores	5.04	345.24	1
23	Mean Difference, Log, Range	0.15	0.05	3
24	Mean Difference, Log, Inter-Decile Range	104.76	2303.92	1
25	Mean Difference, Inverse Hyperbolic Sine, Z-Scores	5.20	348.67	1
26	Mean Difference, Inverse Hyperbolic Sine, Range	0.15	0.05	3
27	Mean Difference, Inverse Hyperbolic Sine, Inter-Decile Range	107.63	2408.57	1

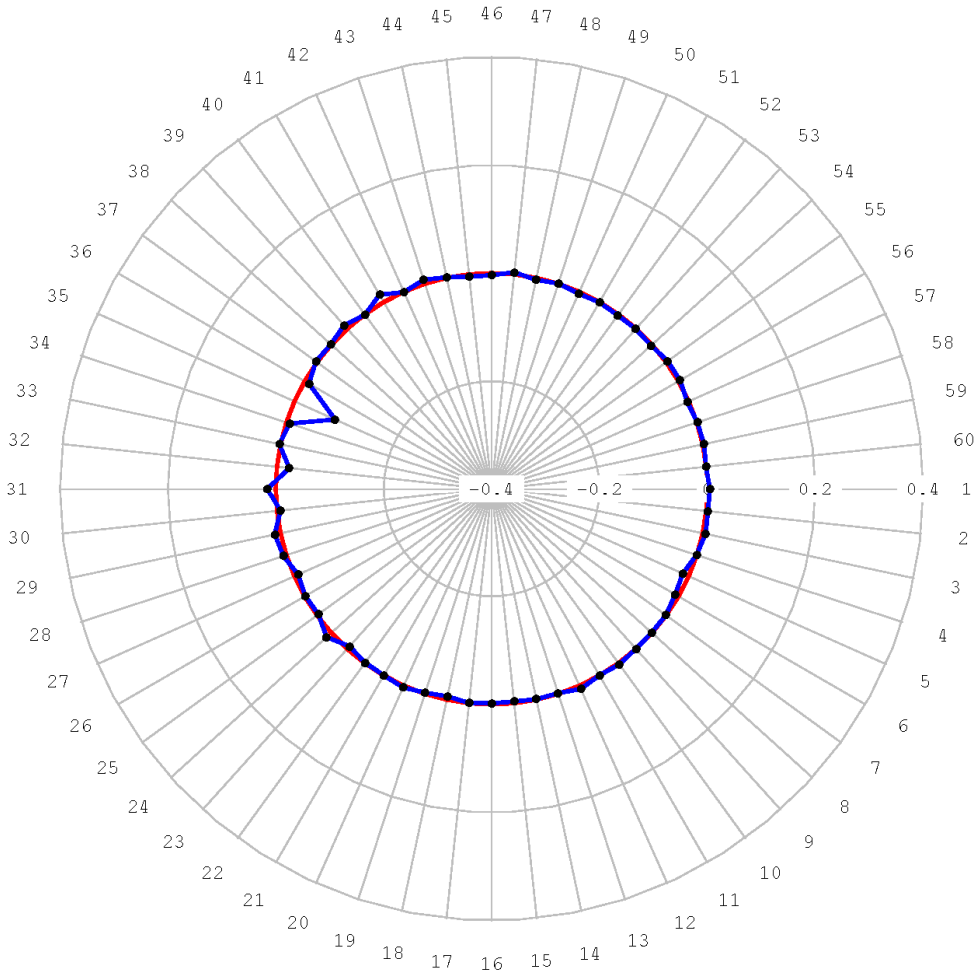


Figure 7.8: Radial plot of a cluster created with the Mean Difference, Box-Cox, Range dataset
(See Table 7.3 for variable names)

The mean squared Euclidean distance (SED) values for each clustering solution, as shown in Table 7.6, were not utilised for the removal of datasets for consideration. Although the values reflect the average homogeneity of each clustering solution to be identified, they do not provide any insight into whether more homogenous clusters ultimately create a general purpose classification that 'looks right' (Mandelbrot, 1982b). It would therefore have been inappropriate to justify the use or removal of a dataset based on the mean SED values alone; particularly as the 2011 OAC Supergroup level requires a certain level of generality to fit all of the characteristics of the UK's population into a small number of groups. The SED value was therefore utilised in the final selection of datasets when the other measures, such as geographic distribution and cluster composition, provided limited distinguishing factors.

7.3.1.2. Optimum cluster numbers

Following application of these three rationales, four datasets remained, all of which had been standardised using the range method. Standardisation on the datasets using the z-scores or inter-decile range techniques had an increased propensity to create substandard cluster solutions. Therefore, the datasets which used these methods were discarded. The identification of the ideal number of clusters in a dataset can be performed using several techniques. Methods, such as silhouette plots (Rousseeuw, 1987) and the gap statistic (Tibshirani et al., 2000) were considered for use, but did not function correctly due to the size of the 2011 OAC dataset. Of those tested, the sole method which functioned as required was the comparison of the total WCSS values for a range of cluster solutions. The most appropriate solution can be defined as that which has a constant total WCSS value and demonstrates the lowest reduction in the total WCSS value for subsequent numbers of clusters (Peeples, 2011). Increasing the number of clusters means greater variance within the dataset is explained. However, the amount of variance explained by each new cluster becomes exponentially smaller, to the point where adding a new cluster adds only limited explanatory power. If plotted, this pattern would produce an 'elbow', and the cluster number where this shape occurred can be considered as the optimum solution. Figure 7.9 shows the results of this technique performed on the four 2011 OAC datasets for 2 to 20 cluster solutions. As no 'elbow' exists in any dataset, there is no obvious number of clusters each dataset could have been divided into.

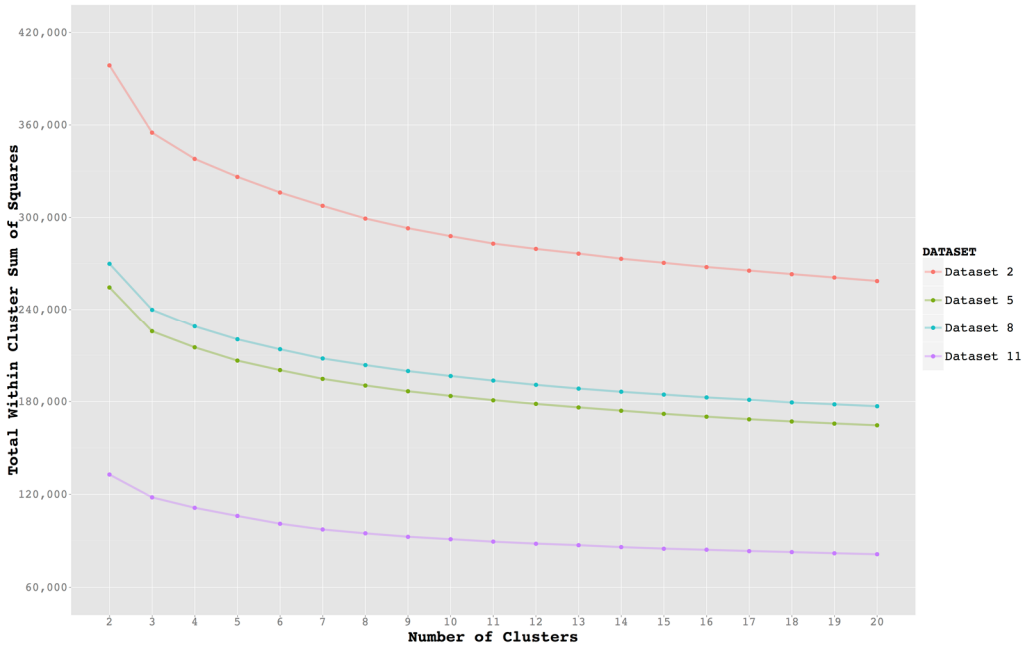


Figure 7.9: WCSS value comparison for 2 to 20 cluster solutions for 4 datasets

As no statistical evidence for the optimum number of clusters for each dataset could be produced, it was necessary to perform exploratory analysis. Each dataset was tested to understand how different numbers of clusters influence the final outcome, and which offered the optimum solution from both a statistical perspective and in terms of which solution looked visually the most representative, or right (Mandelbrot, 1982b). In total 20 unique dataset permutations would have been created by looking at 5 to 9 cluster solutions for the four datasets. To reduce this number, the numbers of clusters considered to form the 2011 OAC Supergroups was reduced to be 6 to 8. These numbers of clusters allowed for flexibility in the selection of the number of clusters which best summarised the characteristics of the UK's population, whilst retaining similarity to the 2001 OAC (as highlighted during the 2011 OAC user engagement discussed in Chapter 4).

Figures 7.10 to 7.12 show the percentage of OAs and SAs assigned to each cluster for solutions with 6, 7 and 8 clusters for each dataset. There are minimal differences between the cluster assignment distributions and therefore there is little to distinguish the four datasets or the different cluster solutions. It was therefore necessary to analyse the composition and geographical distribution of the different clusters solutions.

The geographical distribution across the whole of the UK was analysed by looking at a geographic subset of the 21 different clustering solutions. Figures 7.13 to 7.16 show the geographical distributions of 6, 7 and 8 cluster solutions mapped for each of the four datasets in London, Wolverhampton and Glasgow. There are clear differences between how each dataset impacts the clustering solutions, and the impact of different numbers of clusters. The geographical distribution, in conjunction with the cluster profiles, can be used to summarise the effectiveness of each dataset to create the 'best' clusters from a user perspective; allowing the identification of the optimum number of clusters for each dataset.

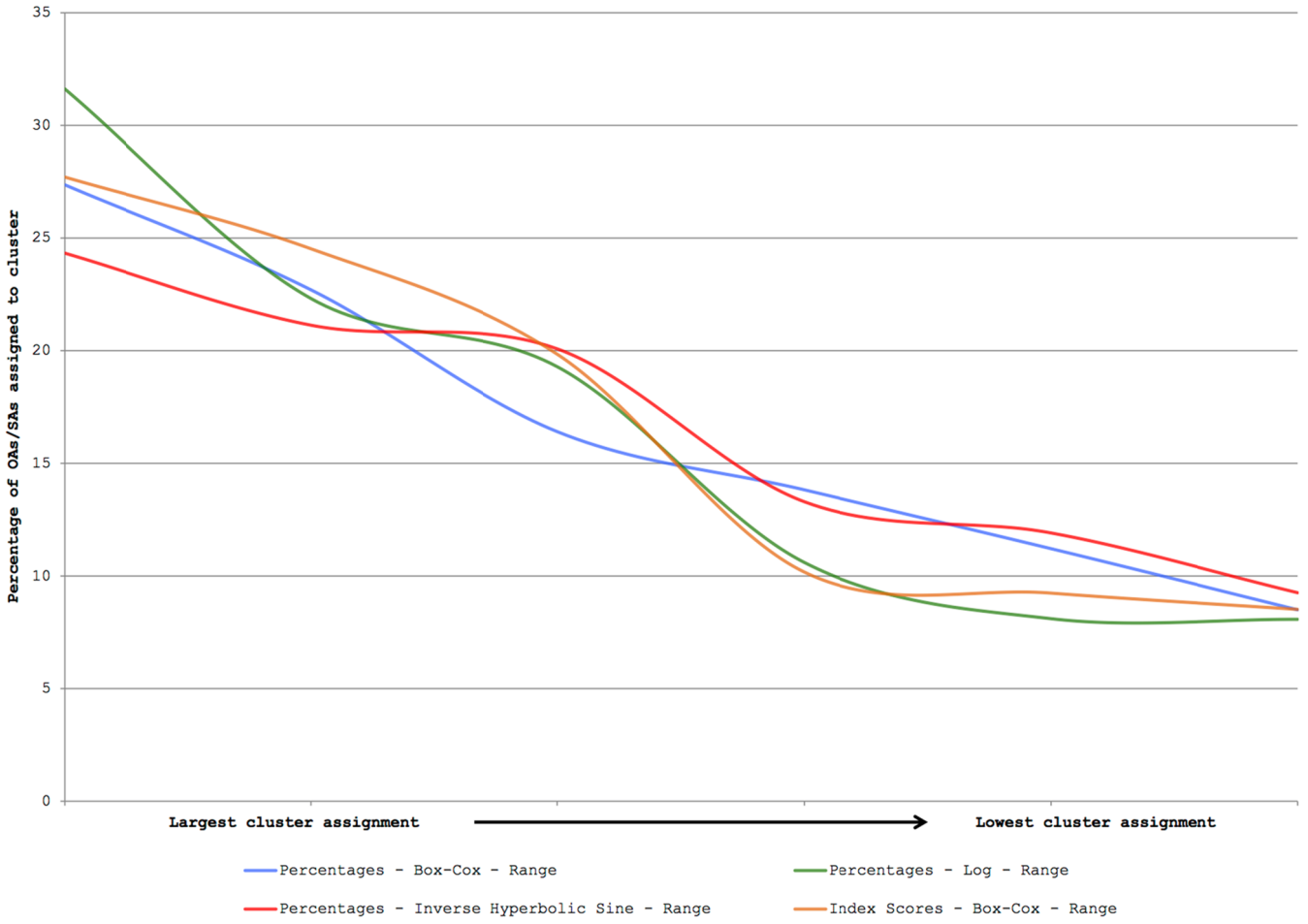


Figure 7.10: Cluster assignment for a 6 cluster solution for 4 datasets

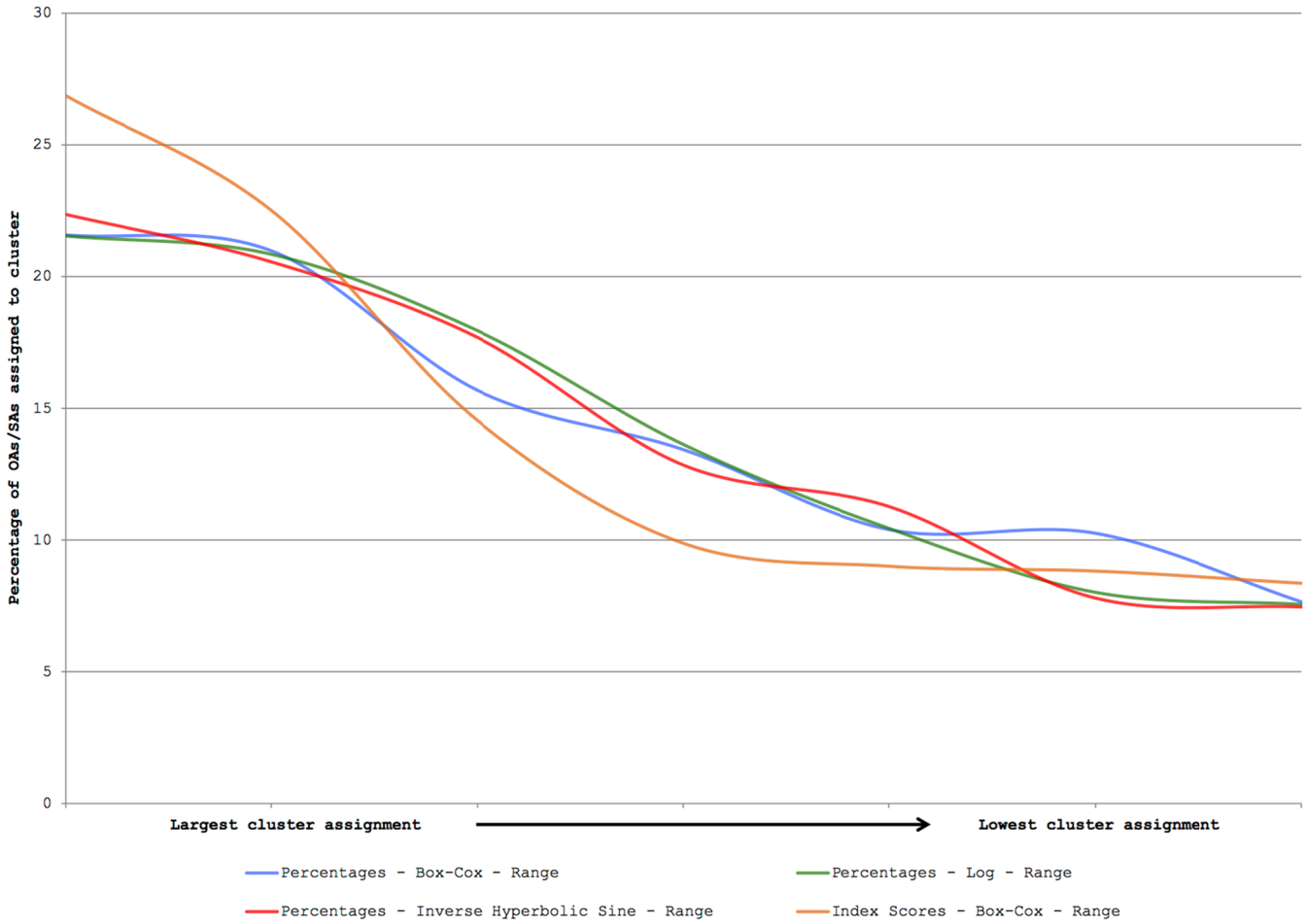


Figure 7.11: Cluster assignment for a 7 cluster solution for 4 datasets

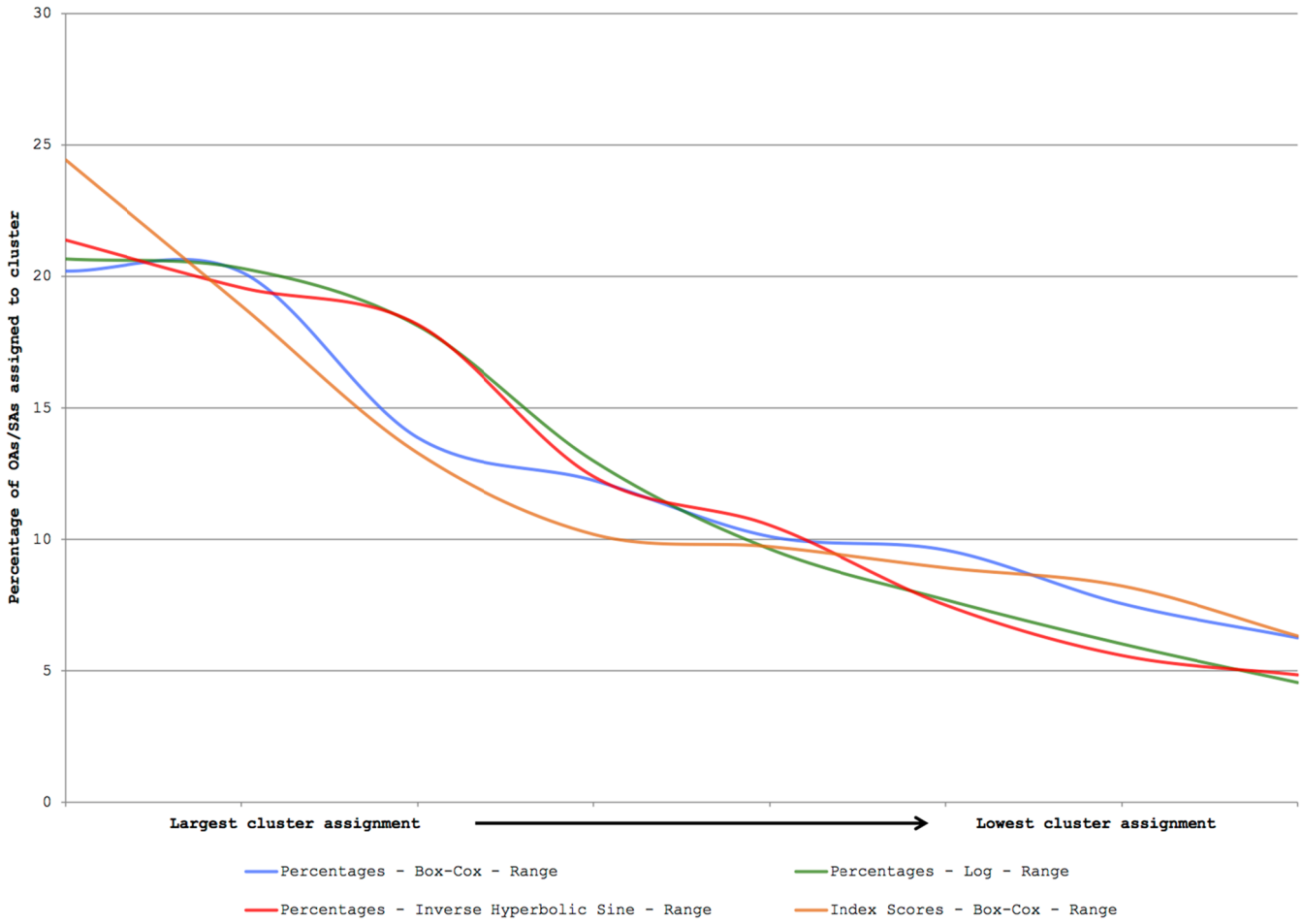


Figure 7.12: Cluster assignment for a 8 cluster solution for 4 datasets

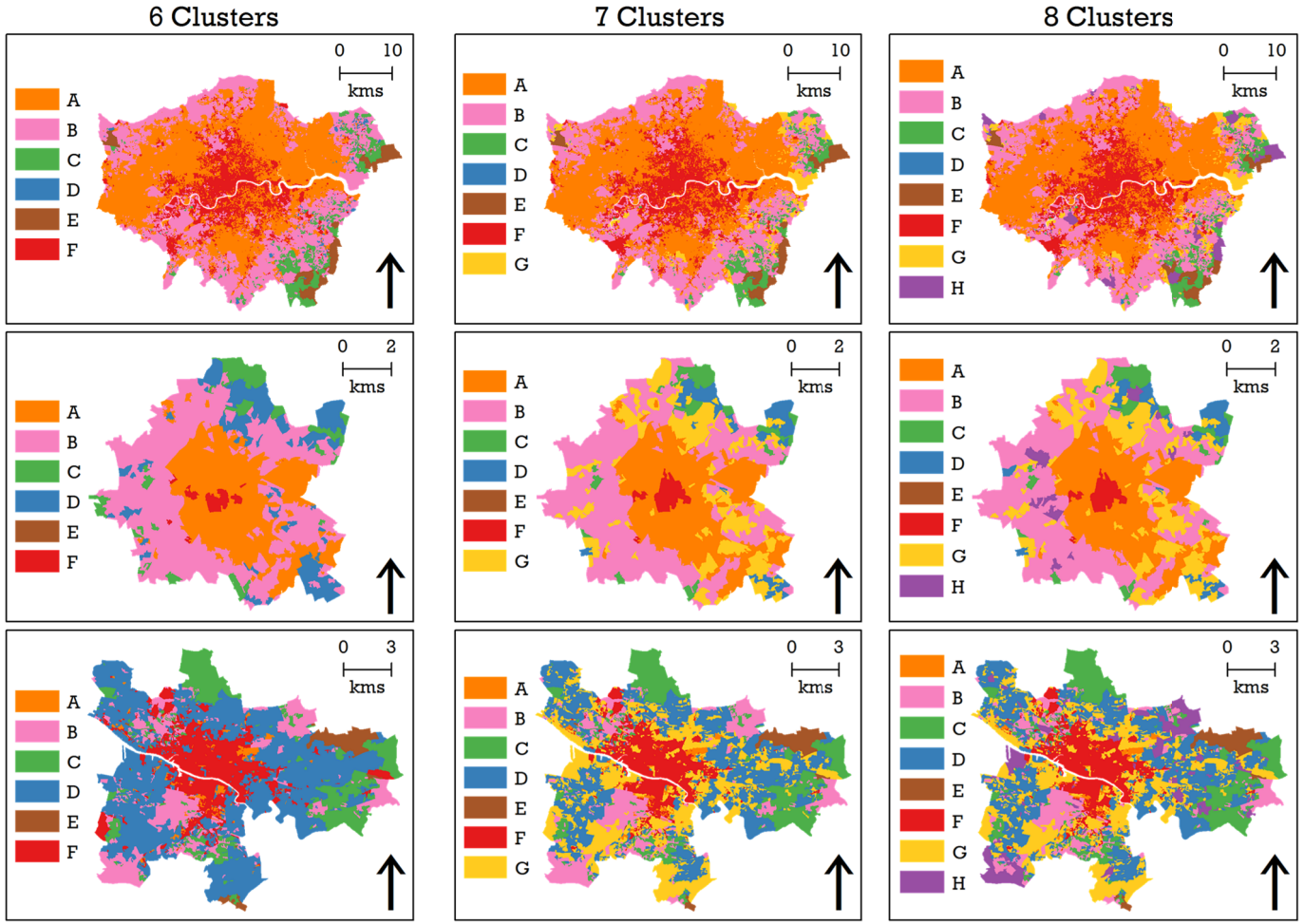


Figure 7.13: Dataset 2 geographic distribution for 6 to 8 cluster solutions in London (top), Wolverhampton (middle) and Glasgow (bottom)

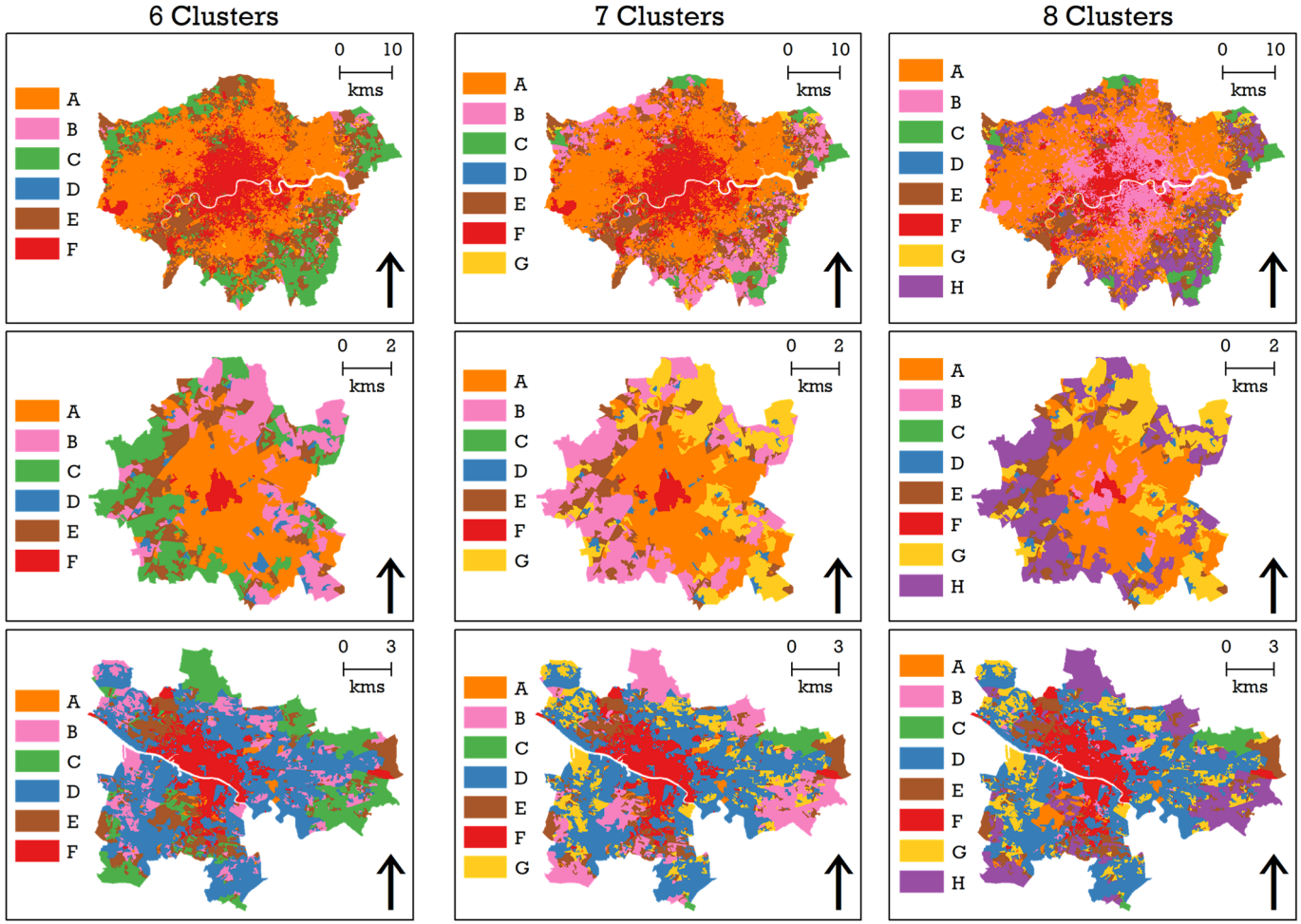


Figure 7.14: Dataset 5 geographic distribution for 6 to 8 cluster solutions in London (top), Wolverhampton (middle) and Glasgow (bottom)

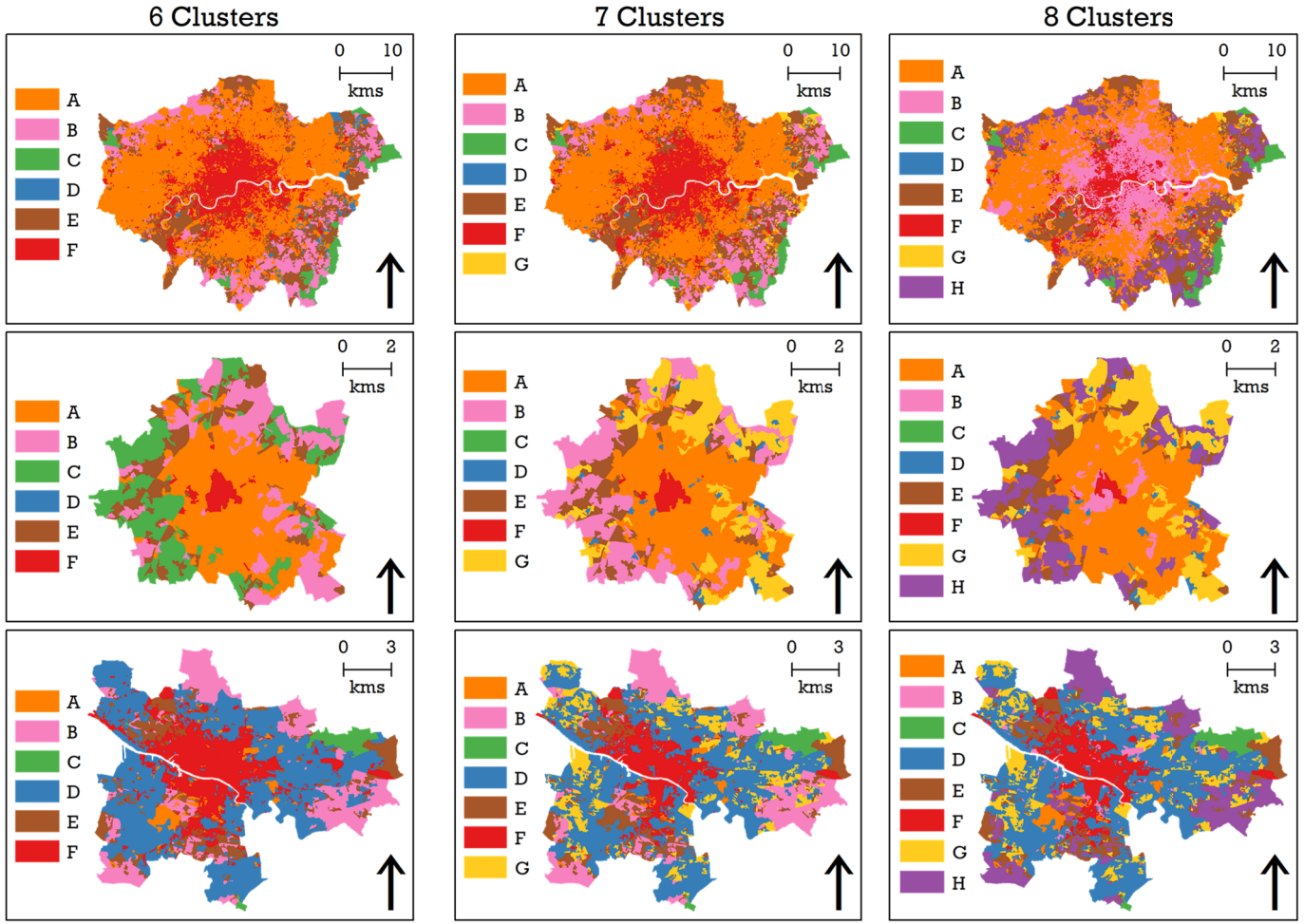


Figure 7.15: Dataset 8 geographic distribution for 6 to 8 cluster solutions in London (top), Wolverhampton (middle) and Glasgow (bottom)

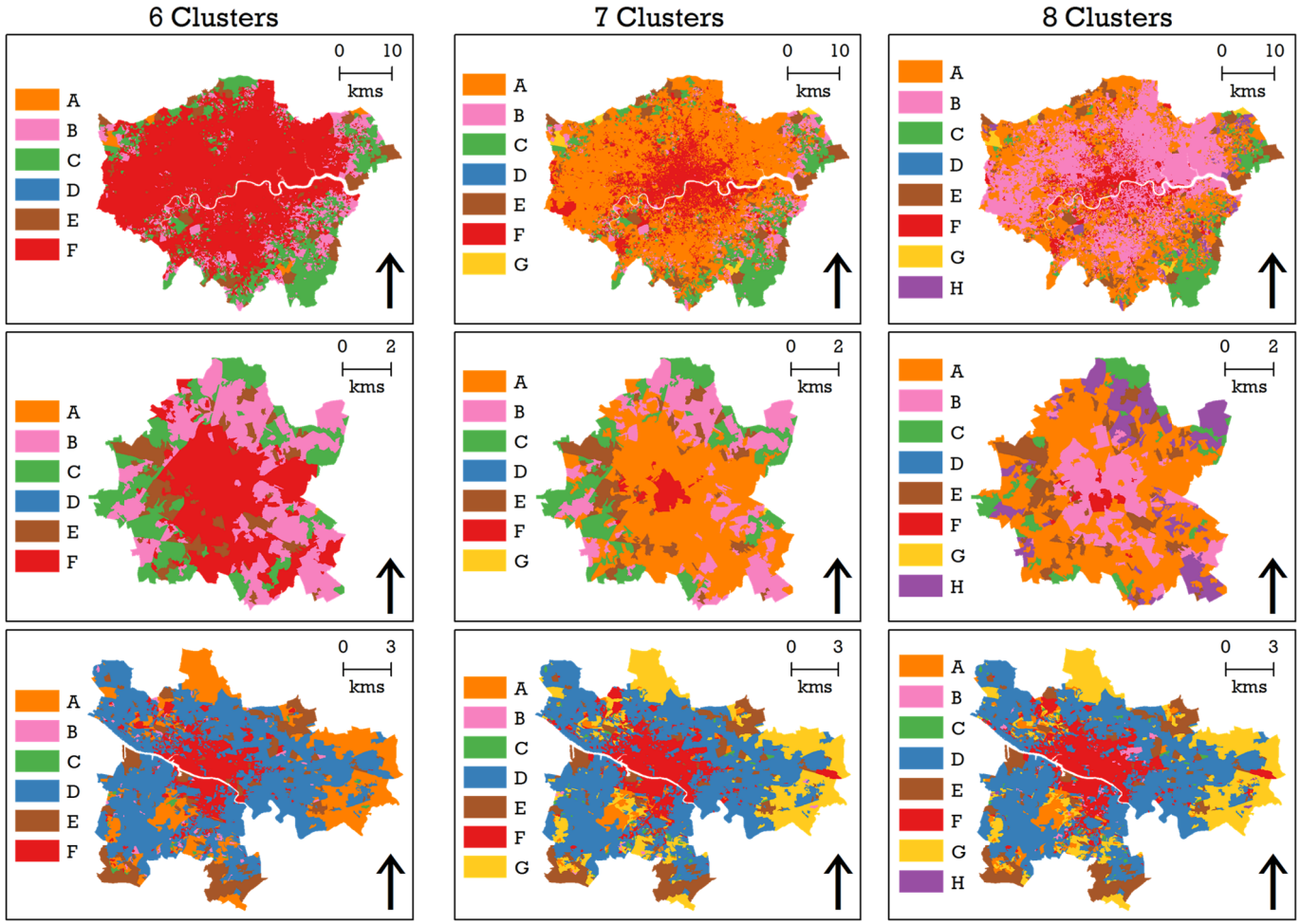


Figure 7.16: Dataset 11 geographic distribution for 6 to 8 cluster solutions in London (top), Wolverhampton (middle) and Glasgow (bottom)

The findings from the four datasets are detailed below, with the geographic distributions of the clusters and the optimum number of groupings for each dataset explained.

Dataset 2 – Percentages, Box-Cox, Range

There is little to distinguish the 7 and 8 cluster solutions in terms of geographic distribution. Both offer better differentiation than the 6 cluster solution, as demonstrated most clearly in Figure 7.13 by Glasgow. However, the geographic variation in the cluster assignments demonstrated are not as good as other datasets, most notably those produced using 'Percentages, Log, Range' and 'Percentages, Inverse Hyperbolic Sine, Range'. In London for example, the 8 cluster solution results in three clusters being assigned to 93.6% of the city's OAs. In addition to this, the differentiation power of the clusters is variable. Present in each of the 6, 7 and 8 cluster solutions is a cluster which represents the national average for the majority of variables, and only loads on 'Persons aged 90 and over' and 'Persons living in a communal establishment' variables, as shown in Figure 7.17. The presence of this cluster in all of the solutions suggests that it reflects genuine characteristics of a portion of the UK's population, but in doing so, creates a very specific Supergroup.

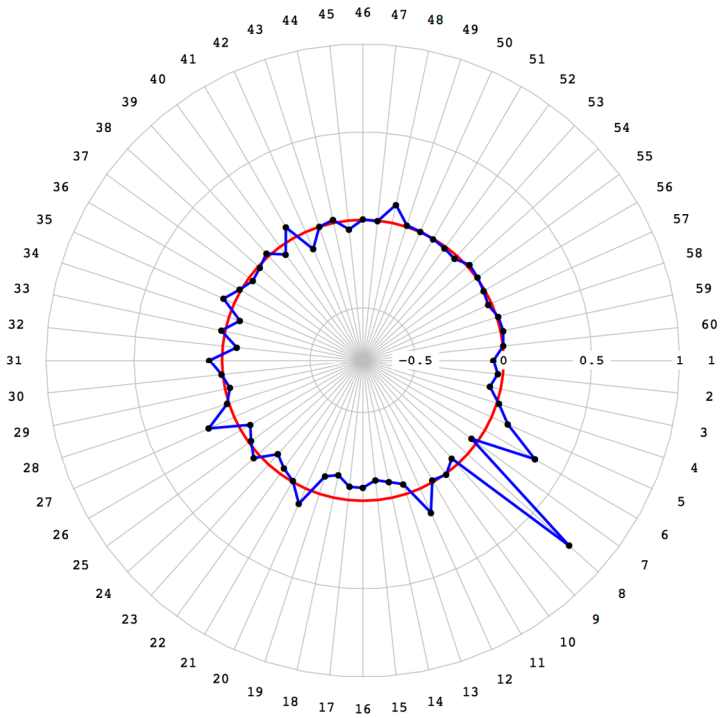


Figure 7.17: Radial plot of a cluster loading on 90+ and communal establishments created with the Percentages, Box-Cox, Range dataset
(See Table 7.3 for variable names)

Dataset 5 – Percentages, Log, Range

The optimum number of clusters for this dataset was 8. A comparison between the 7 and 8 cluster solutions in Figure 7.14 shows that the addition of the extra cluster had a significant impact on London. The 7 cluster solution resulted in the representation of 81% of OAs in London by two clusters, with the 8 cluster solution this value dropped to 67%. The increased geographic variation in the distributions offered by the 8 cluster solution addressed concerns expressed in the 2011 OAC user engagement regarding how relatively poorly London was represented by the 2001 OAC where 77.5% of London OAs were clustered into two groups. The marked improvement offered by the 8 cluster solution was also seen in the cluster profiles, which offered good variation between the groups. Notably the two clusters found in inner London had distinctive characteristics.

Dataset 8 – Percentages, Inverse Hyperbolic Sine, Range

This dataset is similar to the 'Percentages, Log, Range' dataset in terms of the geographic distribution of clusters and their profiles for the 7 and 8 cluster solutions (see Figure 7.15). This similarity is likely to be a result of the parallels between the Inverse Hyperbolic Sine (IHS) and log transformation techniques. However, as noted by Burbidge et al. (1988), the IHS method is better at handling datasets which contain a large number of zeros. The final 60 variables contain a varying number of zero values, and the way in which the IHS transformation method handles the variables where they are more prevalent, produces the slightly different results to the 'Percentages, Log, Range' dataset. Although both datasets suggest an 8 cluster solution is optimum, the 'Percentages, Inverse Hyperbolic Sine, Range' dataset provides an alternative solution that incorporates a transformation method better designed to handle the type of data found in the 2011 UK Census.

Dataset 11 – Index Scores, Box-Cox, Range

This dataset was unstable in comparison to those which underwent the conversation of their raw data to percentages. The geographic distribution of clusters for the different solutions, as shown in Figure 7.16, shows large variations between them. These variations are most apparent in London, and to a lesser extent in Wolverhampton. Glasgow, in comparison, demonstrates cluster stability between the three solutions. This wide variation means that 8 clusters can be considered to be the optimum number, particularly in London. The 8 cluster solution offered by this dataset is however not as good in either the geographical distribution of the clusters, or their composition, when compared to the Percentages, Log, Range' or 'Percentages, Inverse Hyperbolic Sine,

Range' datasets. This dataset for example results in 76% of OAs in London being represented by two clusters in the 8 cluster solution, compared to the 70% offered by the 'Percentages, Inverse Hyperbolic Sine, Range' dataset.

Analysis of the four datasets suggested that in each case an 8 cluster solution produced the best result. As a result, 8 clusters were chosen as the number of Supergroups for the 2011 OAC. The selection of the best dataset was not as straightforward however, as the dataset with the lowest mean SED value did not have the best geographical dispersion of clusters. As discussed previously, the most desirable characteristic for the 2011 OAC Supergroups was their ability to differentiate between different areas of the UK, even if this resulted in a reduction in homogeneity. The 'Percentages, Inverse Hyperbolic Sine, Range' dataset was therefore selected to create the 2011 OAC. This dataset created clusters with good geographical distribution that resembled known variations in the UK population and therefore can be considered to 'look right' (Mandelbrot, 1982b). The cluster profiles had sufficient differentiation power that unique names and descriptions could be formed. Additionally, the 'Percentages, Inverse Hyperbolic Sine, Range' dataset was preferable because of the advantages offered by the IHS transformation method, in particular how it handles zero values. This meant modification of the 2011 UK Census data was reduced by not requiring a constant to be added that other transformation techniques would have needed to function correctly. The subjective nature of this decision means that others could justify an alternative selection, particularly if the main criteria for selecting a dataset were based on producing as homogenous clusters as possible.

7.3.2. Creating a hierarchical classification

The choice of the 'Percentages, Inverse Hyperbolic Sine, Range' dataset and an 8 cluster solution to form the 2011 OAC Supergroups had a direct impact on the formation of the lower tiers of the classification. As the non-hierarchical k-means clustering algorithm was used, it was necessary to manually cluster the data to create the Groups and Subgroups. The choice of the same top-down method as used by Vickers et al. (2005) (as discussed in Section 6.8.4) meant the k-means algorithm was run 8 times on the Supergroup dataset; each time on a dataset comprising of only OAs or SAs assigned to a different Supergroup. This process was subsequently repeated on each of the Group datasets to form the Subgroups of the 2011 OAC. Also discussed in Section 6.8.4 was the decision to consider 2 to 4 clusters to make up the Groups derived from the Supergroups

and the Subgroups derived from the Groups. The final hierarchical structure of the 2011 OAC contained an additional Supergroup in comparison to the 2001 OAC. Limiting the number of clusters per Group or Subgroup to a maximum of 4 in the 2011 OAC avoided a large increase in the total number of Groups and Subgroups when compared to the previous classification.

The decision on the exact number of Groups and Subgroups was based on similar criteria used to select the final number of Supergroups. The geographic variation of the clusters was not considered to be as important. The focus was instead on ensuring that cluster solutions did not lead to the creation of micro clusters and that the clusters offered good differentiation power. Table 7.7 identifies the variance between cluster sizes and the mean SED value for each Group solution. Values closer to 0 indicate a greater disparity between the number of OAs and SAs assigned to the largest and smallest clusters, suggesting the presence of micro clusters. The 'Mean SED' column identifies the overall cluster homogeneity of each solution. A smaller number indicates a greater level of cluster homogeneity. This value should decrease with an increase in cluster numbers, meaning that the cluster solution with the smallest value was not guaranteed to be used as this was invariably the solution with 4 clusters.

As with the optimum dataset and optimum number of clusters for the 2011 OAC Supergroups, decisions regarding cluster solutions were only based on the mean SED values if no other justification could be found. These values were however useful in exploring how the datasets for each Supergroup were clustered differently, with some of the solutions for the derived Groups creating more homogenous clusters than others.
Table 7.7 provided a statistical method of determining the optimum number of Groups to form the 2011 OAC. Looking solely at Table 7.7, the number of clusters derived from Supergroup 1 should have been two. This solution offered clusters that were the most evenly sized and as homogenous as the three cluster alternative. However, this would not have taken into account the differentiation offered by the clusters created from each solution. As such, the cluster profiles of each solution needed to be examined.

Table 7.7: Cluster size variance and mean SED of potential 2011 OAC Groups (final selection highlighted in green)

	2 Clusters		3 Clusters		4 Clusters	
	Cluster size variance	Mean SED	Cluster size variance	Mean SED	Cluster size variance	Mean SED
Supergroup 1	0.75	0.68	0.25	0.68	0.35	0.65
Supergroup 2	0.45	1.39	0.41	1.27	0.60	1.19
Supergroup 3	0.61	0.93	0.88	0.85	0.42	0.88
Supergroup 4	0.67	0.89	0.56	0.83	0.39	0.80
Supergroup 5	0.83	0.81	0.29	0.80	0.34	0.77
Supergroup 6	0.73	0.70	0.40	0.68	0.44	0.65
Supergroup 7	0.35	1.30	0.50	1.17	0.37	1.17
Supergroup 8	0.95	0.68	0.94	0.65	0.71	0.63

Figures 7.18 and 7.19 shows an example of a cluster solution that offers good differentiation power and one that does not. The clusters in Figure 7.18 show a number of variables that deviate away from the mean (depicted as the red line), allowing unique names and descriptions to be formed. Conversely, Figure 7.19 shows three out of the four clusters have a majority of variables that do not deviate significantly from the mean, and as such offer only limited differentiation power. Consequently, a solution demonstrating cluster profiles like that in Figure 7.19 would not be selected. Taking into account the cluster profile and results in Table 7.7, a three cluster solution was selected for Supergroup 8. This solution offered the best differentiation from the options available and would therefore be the most useful to users of the 2011 OAC. In total, 26 Groups were created from the Supergroups. The final selection is highlighted in green in Table 7.7. The 76 Subgroups were selected utilising the same methodology and rationale as in the Group selection. The solutions selected are highlighted in green in Table 7.8.

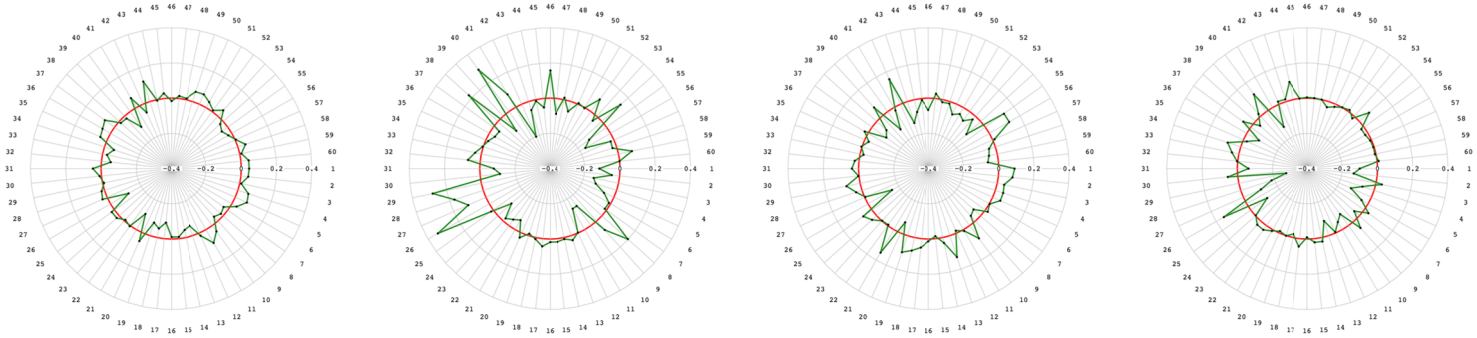


Figure 7.18: Radial plots showing clusters with good differentiation

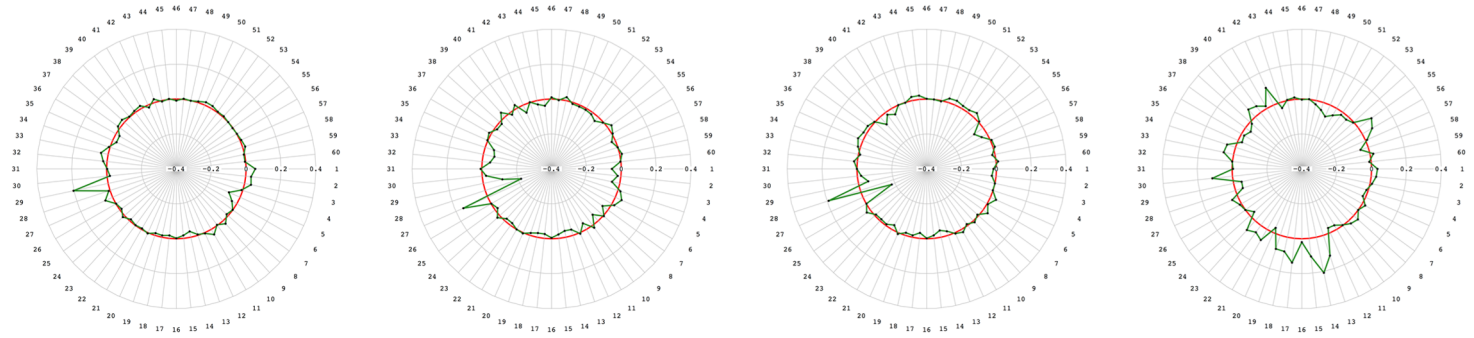


Figure 7.19: Radial plots showing clusters with poor differentiation

Table 7.8: Cluster size variance and mean SED of potential 2011 OAC Subgroups (final selection highlighted in green)

	2 Clusters		3 Clusters		4 Clusters	
	Cluster size variance	Mean SED	Cluster size variance	Mean SED	Cluster size variance	Mean SED
Group 1a	0.75	0.65	0.79	0.62	0.51	0.62
Group 1b	0.96	0.54	0.68	0.53	0.70	0.52
Group 1c	0.66	0.74	0.83	0.70	0.59	0.70
Group 2a	0.35	1.15	0.31	1.03	0.25	0.98
Group 2b	0.84	1.26	0.22	1.19	0.28	1.16
Group 2c	0.80	1.21	0.70	1.16	0.50	1.10
Group 2d	0.68	0.77	0.72	0.73	0.67	0.70
Group 3a	0.79	0.70	0.31	0.72	0.30	0.68
Group 3b	0.41	0.68	0.48	0.64	0.46	0.63
Group 3c	0.52	1.31	0.65	1.22	0.44	1.21
Group 3d	0.60	0.62	0.77	0.58	0.45	0.58
Group 4a	0.76	0.73	0.66	0.69	0.67	0.67
Group 4b	0.74	0.79	0.67	0.75	0.66	0.73
Group 4c	0.79	0.75	0.78	0.70	0.73	0.67
Group 5a	0.76	0.71	0.79	0.67	0.64	0.66
Group 5b	0.35	0.81	0.49	0.79	0.52	0.75
Group 6a	0.50	0.73	0.46	0.71	0.48	0.68
Group 6b	0.87	0.57	0.32	0.57	0.35	0.55
Group 7a	0.90	0.74	0.54	0.73	0.22	0.72
Group 7b	0.90	1.34	0.68	1.31	0.47	1.29
Group 7c	0.82	0.85	0.86	0.82	0.65	0.82
Group 7d	0.26	1.74	0.28	1.52	0.20	1.56
Group 8a	0.78	0.56	0.50	0.54	0.37	0.54
Group 8b	0.89	0.64	0.59	0.64	0.79	0.60
Group 8c	0.57	0.60	0.58	0.58	0.14	0.56
Group 8d	0.90	0.57	0.56	0.55	0.54	0.53

In total 110 clusters were selected to create the 2011 OAC: 8 Supergroups, 26 Groups and 76 Subgroups. The breakdown of how each of these clusters formed the 2011 OAC hierarchy is shown in Table 7.9, along with the total number of UK OAs and SAs that were assigned to each group. The Supergroups range in the number of OAs and SAs assigned to them from 20.6% for Supergroup 6 to 5.1% for Supergroup 3. Although this range is larger than that of the 2001 OAC Supergroups, this greater number of smaller clusters is useful in increasing the discriminatory power of the classification, especially in urban areas. The size of each Group and Subgroup is influenced by the relative size (in terms of OAs and SAs assigned) of their parent Supergroup along with the number of clusters decided upon. This variance results in Group 6b representing 11.6% of the UK's OAs and SAs, and Group 3c a much smaller 0.7%. Adding or subtracting an extra Group would not have changed these values sufficiently to provide uniformity, nor would it have been an appropriate action due to the adverse impact this would have had on the cluster compositions.

A similar discrepancy exists between the Subgroups, with the highest and lowest OA and SA assignments across the UK ranging from 4.2% to 0.1%. Although the differences in how many OAs and SAs are assigned to each Group and Subgroup cannot be ignored, the significance of the variation is negligible. The use of the k-means clustering algorithm in the creation of the 2011 OAC on the 8 separate datasets in the formation of the Groups, and subsequently on the 26 separate datasets to form the Subgroups means that the 2011 OAC is in effect 35 different classifications combined to make a three-tiered hierarchy. This methodology ensures that each OA and SA is assigned to one Supergroup, Group and Subgroup, but as a result not all clusters are directly connected. For example, as Groups 2a and 2b share the same Supergroup they are related, however Groups 2b and 4b are not, even though they occupy the same level of hierarchy in the classification and may be geographically next to each other in some locations. The focus of creating clusters with the best differentiation power, combined with the methodology that meant Subgroups nested within Groups, and Groups within Supergroups means variances between the number of OAs and SAs assigned to each cluster was an expected consequence.

Table 7.9: The hierarchical structure of the 2011 OAC

Supergroup	Total UK OAs & SAs	Group	Total UK OAs & SAs	Subgroup	Total UK OAs & SAs
Supergroup 1	11.75% (27300)	Group 1a	4.38% (10164)	Subgroup 1a1	2454 (1.06%)
				Subgroup 1a2	3086 (1.33%)
				Subgroup 1a3	3043 (1.31%)
				Subgroup 1a4	1581 (0.68%)
		Group 1b	5.89% (13683)	Subgroup 1b1	5272 (2.27%)
				Subgroup 1b2	4811 (2.07%)
				Subgroup 1b3	3600 (1.55%)
		Group 1c	1.49% (3453)	Subgroup 1c1	1030 (0.44%)
				Subgroup 1c2	1183 (0.51%)
				Subgroup 1c3	1240 (0.53%)
Supergroup 2	5.65% (13125)	Group 2a	1.10% (2561)	Subgroup 2a1	467 (0.20%)
				Subgroup 2a2	574 (0.25%)
				Subgroup 2a3	1520 (0.65%)
		Group 2b	1.15% (2670)	Subgroup 2b1	1449 (0.62%)
				Subgroup 2b2	1221 (0.53%)
		Group 2c	1.84% (4267)	Subgroup 2c1	1683 (0.72%)
				Subgroup 2c2	1182 (0.51%)
				Subgroup 2c3	1402 (0.60%)
		Group 2d	1.56% (3627)	Subgroup 2d1	1002 (0.43%)
				Subgroup 2d2	1386 (0.60%)
				Subgroup 2d3	1239 (0.53%)
Supergroup 3	5.10% (11849)	Group 3a	1.51% (3509)	Subgroup 3a1	1958 (0.84%)
				Subgroup 3a2	1551 (0.67%)
		Group 3b	1.28% (2963)	Subgroup 3b1	1382 (0.59%)
				Subgroup 3b2	660 (0.28%)
				Subgroup 3b3	921 (0.40%)
		Group 3c	0.68% (1586)	Subgroup 3c1	1043 (0.45%)
				Subgroup 3c2	543 (0.23%)
		Group 3d	1.63% (3791)	Subgroup 3d1	1094 (0.47%)
				Subgroup 3d2	1419 (0.61%)
				Subgroup 3d3	1278 (0.55%)
Supergroup 4	10.12% (23502)	Group 4a	4.71% (10942)	Subgroup 4a1	4488 (1.93%)
				Subgroup 4a2	3512 (1.51%)
				Subgroup 4a3	2942 (1.27%)
		Group 4b	2.65% (6146)	Subgroup 4b1	3540 (1.52%)
				Subgroup 4b2	2606 (1.12%)
		Group 4c	2.76% (6414)	Subgroup 4c1	2322 (1.00%)
				Subgroup 4c2	1803 (0.78%)
				Subgroup 4c3	2289 (0.99%)
Supergroup 5	16.66% (38697)	Group 5a	9.09% (21124)	Subgroup 5a1	8046 (3.46%)
				Subgroup 5a2	6378 (2.75%)
				Subgroup 5a3	6700 (2.88%)
		Group 5b	7.56% (17573)	Subgroup 5b1	4961 (2.14%)
				Subgroup 5b2	4130 (1.78%)
				Subgroup 5b3	8482 (3.65%)

Supergroup	Total UK OAs & SAs	Group	Total UK OAs & SAs	Subgroup	Total UK OAs & SAs
Supergroup 6	20.17% (46850)	Group 6a	8.52% (19801)	Subgroup 6a1	3409 (1.47%)
				Subgroup 6a2	4353 (1.87%)
				Subgroup 6a3	7174 (3.09%)
				Subgroup 6a4	4865 (2.09%)
		Group 6b	11.64% (27049)	Subgroup 6b1	3434 (1.48%)
				Subgroup 6b2	9860 (4.24%)
				Subgroup 6b3	8178 (3.52%)
				Subgroup 6b4	5577 (2.40%)
Supergroup 7	11.68% (27135)	Group 7a	4.04% (9389)	Subgroup 7a1	2170 (0.93%)
				Subgroup 7a2	3214 (1.38%)
				Subgroup 7a3	4005 (1.72%)
		Group 7b	2.05% (4752)	Subgroup 7b1	1344 (0.58%)
				Subgroup 7b2	1428 (0.61%)
				Subgroup 7b3	1980 (0.85%)
		Group 7c	4.09% (9508)	Subgroup 7c1	3382 (1.46%)
				Subgroup 7c2	3225 (1.39%)
				Subgroup 7c3	2901 (1.25%)
		Group 7d	1.50% (3486)	Subgroup 7d1	1526 (0.66%)
				Subgroup 7d2	909 (0.39%)
				Subgroup 7d3	739 (0.32%)
				Subgroup 7d4	312 (0.13%)
Supergroup 8	18.87% (43838)	Group 8a	4.94% (11474)	Subgroup 8a1	6448 (2.78%)
				Subgroup 8a2	5026 (2.16%)
		Group 8b	3.93% (9134)	Subgroup 8b1	4289 (1.85%)
				Subgroup 8b2	4845 (2.09%)
		Group 8c	5.51% (12789)	Subgroup 8c1	5527 (2.38%)
				Subgroup 8c2	3209 (1.38%)
				Subgroup 8c3	4053 (1.74%)
		Group 8d	4.49% (10441)	Subgroup 8d1	3527 (1.52%)
				Subgroup 8d2	4435 (1.91%)
				Subgroup 8d3	2479 (1.07%)

Although the creation of Groups and Subgroups that had a more even distribution in their assignment to OAs and SAs would have been feasible, doing so would have meant creating sub-optimum clusters with reduced differentiation power. As such, both the finalised total number of clusters that form the 2011 OAC and the variation in the assignments to OAs and SAs can be considered the optimum result in creating a general-purpose geodemographic classification of the UK using the 2011 UK Census.

7.4. Outputs

An essential part of building the 2011 OAC was the creation of a number of key outputs to help better understand the results of the clustering process. These outputs can be broadly summarised into two categories: descriptive and visual. The descriptive outputs provide the clusters with enhanced meaning. The assignment of names and descriptions, or pen portraits, to each of the 110 clusters enables users to interpret the results of the clustering process in terms of the characteristics of each cluster, which the output values of the k-means clustering algorithm alone does not allow. The visual outputs from the classification compliment the descriptions, in both understanding the variance between the make-up of each cluster and the geographic distribution of the Supergroups, Groups and Subgroups. Although development of the underlying processes and procedures is fundamental for the creation of the classification, it is these outputs that will have the greatest overall impact, as they will influence how users use and interpret the 2011 OAC.

7.4.1. Clustering outputs

The naming and description of clusters provides a user focused geodemographic classification. The 2011 OAC uses two methods to interpret the initial output cluster information for the formation of names and descriptors: radial plots and bar graphs. The radial plots and bar graphs for the 2011 OAC Supergroups are shown in Figures 7.20 and 7.21 respectively, the names for each cluster correspond to those assigned based on the average characteristics of each Supergroup and are discussed in detail in Section 7.4.2.

Radial plots were previously used by Vickers (2006), where the red line represents the mean value for each variable. In the case of the Supergroups this is the UK mean, and for the Groups and Subgroups there is one mean in relation to the parent Supergroup or Group and another for the UK. The blue line represents how far each variable in a particular cluster deviates away from the mean. This allows the average characteristics of individual groups to be identified, where the variations in the concentration of the 60 2011 OAC variables make each cluster unique. A criticism of this method however is that it implies variables are linked, by joining the variables together, where no such relationship exists. For example, the values between variable k006 (Persons ages 90 and over) and variable k007 (Number of persons per hectare) in the 2011 OAC Supergroup radial plots are linked together, despite the two variables measuring different aspects of the population and they do not share the same denominator. It is these linkages that give the radial plots their distinctive look, but can be misleading as a result.

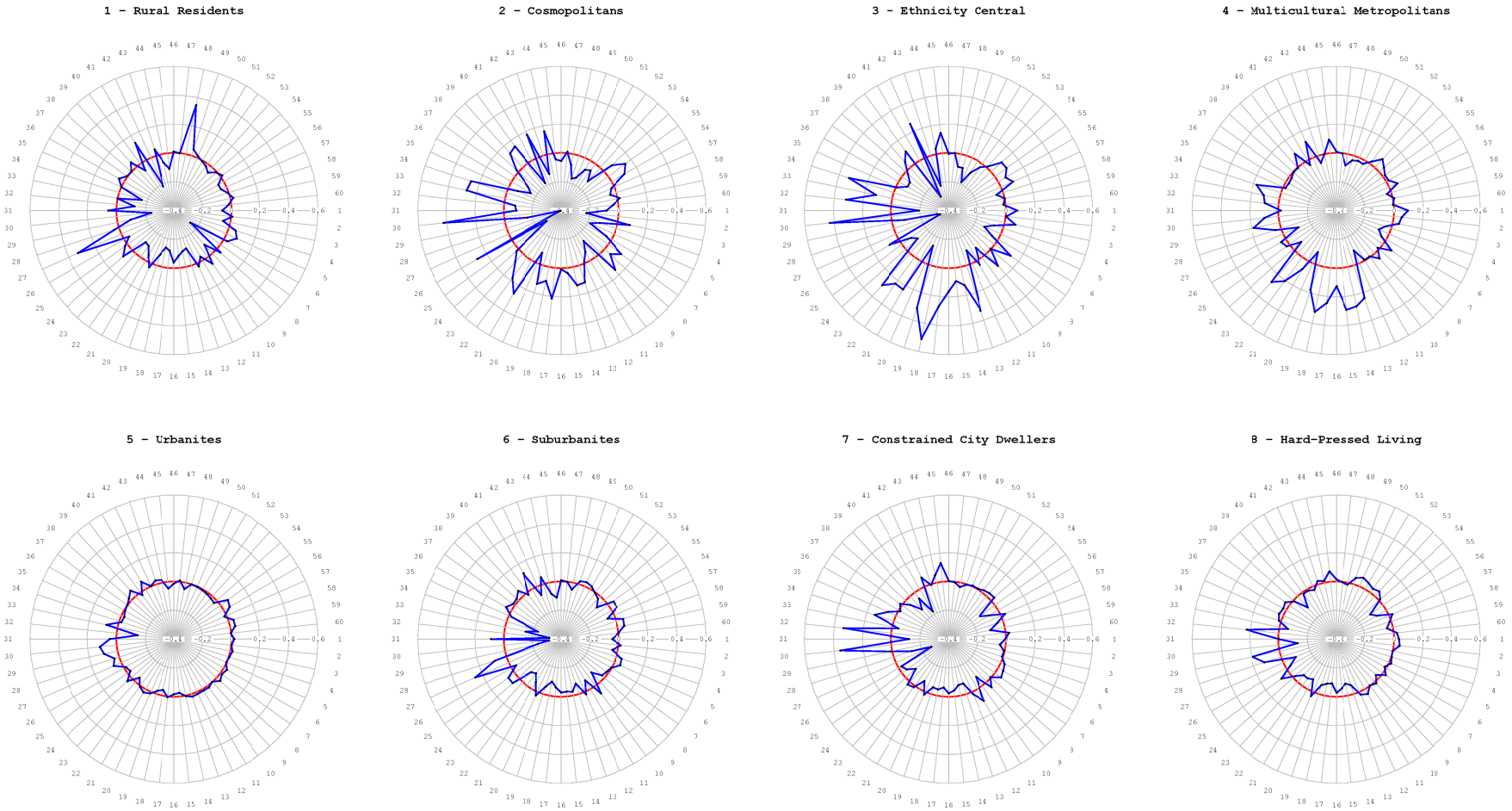


Figure 7.20: Radial plots of the 2011 OAC Supergroups

(See Table 7.3 for variable names)

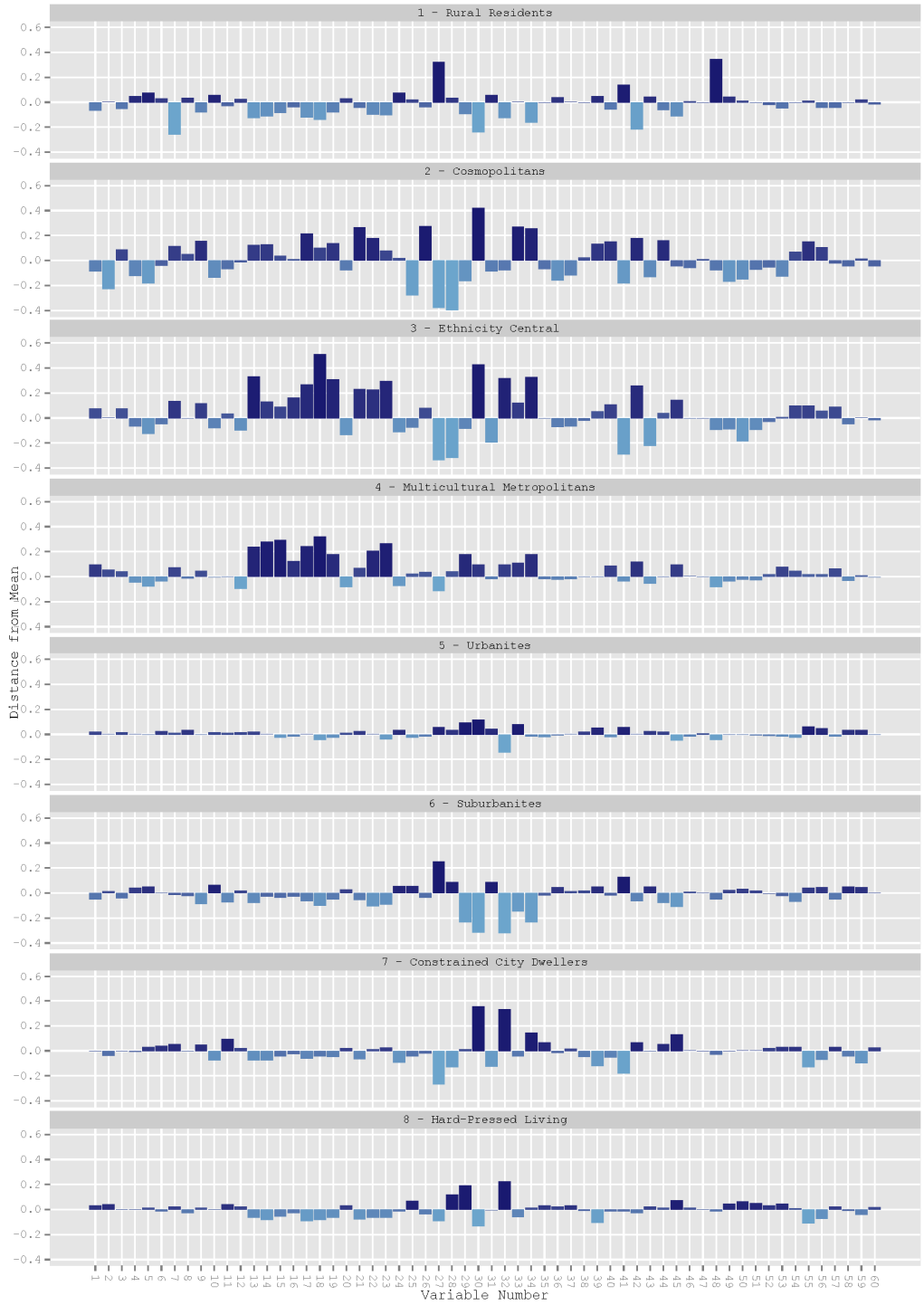


Figure 7.21: Bar graphs of the 2011 OAC Supergroups

(See Table 7.3 for variable names)

The bar graphs in Figure 7.21 also represent the distance away from the mean value for each variable, but they do so in a way that avoids any misleading interpretations. In addition, they are ranked according to their scores, the darkest colour representing the variable with the positive value furthest from the mean and the lightest representing the negative value furthest from the mean. This differentiation in colour allows the identification of the variables which deviate the most from the mean in each cluster solution. The creation of both the radial plots and bar graphs can be considered to be key outputs of the 2011 OAC. They formed the basis on which the cluster names and descriptions were formed, and provide a user with an easy visual reference as to the differing compositions of the clusters.

7.4.2. Naming the clusters

The naming of geodemographic classifications clusters can be a contentious issue and the validity of each name is open to debate. Commercial classifications such as Mosaic (Experian, 2010) and Acorn (CACI, 2013b) only allow for limited re-interpretation of the clusters and subsequent re-naming by users as limited factual information is provided about the formation of the groups. Users must therefore place faith that clusters with names like 'Clocking Off' and 'Urban Cool' provides a representative description of the individuals and households that live in those assigned areas. To make the 2011 OAC as transparent as possible, all output materials used in the naming of the clusters, will be made available once the classification is officially launched by the Office for National Statistics (ONS). This will provide users with the flexibility to re-interpret the underlying data to construct names they believe to be more reflective of the resident population. This flexibility is one of the strengths of an open geodemographic classification, and sets the 2011 OAC apart from the commercial products.

As discussed in Section 4.3.4, an output of the 2011 OAC user engagement was the desire to name all of the Subgroups. This is in line with several commercial classifications, such as Acorn, the latest version of which includes names for all 88 clusters (CACI, 2013b). It was therefore decided to name all clusters of the 2011 OAC. In order to do this, a set of basic principles were followed. Firstly, the names could not be offensive and had to avoid stereotyping. Secondly, each name had to strike a balance between being too descriptive, thereby excluding portions of the resident population, and not so generic that the clusters could not be differentiated from each other. Thirdly, they had to treat the population the same. A name such as 'Black Deprivation' by itself would have been

considered unacceptable. If however it formed a set of Subgroups with 'White Deprivation' and 'Asian Deprivation', then it would have been more likely to be accepted as in this case no single ethnic group has been highlighted as deprived when others are not. Lastly, if possible, the names could not have been used in any previous geodemographic classification.

The primary tools utilised in the naming of the clusters were the radial plots and bar graphs. As discussed in Section 7.4.1, the radial plot for each cluster at the Supergroup level depict deviations of the mean values from the national mean, and for the Groups and Subgroups this is the mean of their respective parent Supergroup and Group. Clusters containing variables demonstrating extreme values, either below or above the mean, were easier to name than those containing variables all with values close to the national or cluster mean because they contained distinct population characteristics. Certain aspects of clusters, such as their geographic distribution, could not be derived from the radial plots or bar graphs. It was therefore necessary to map the clusters to allow the spatial variance of the clusters to be incorporated into the names. Final names could therefore incorporate the resident population, the built environment and geographic location, consequently making it easier to distinguishing clusters, particularly in urban areas. The mapping of clusters also allowed for internal validation, where names were checked against local knowledge of areas. For example, clusters that were believed to have large student populations, and named accordingly, matched areas in Southampton and Bristol known for their large residential student population.

Internal validation can only provide limited reassurance of the accuracy of the constructed names. The true test will be when users apply local knowledge to assess the validity and suitability of names within a geodemographic classification. It is likely that the accuracy of assigned names will decrease at the lower levels of the classification as the clusters at the Subgroup level represent increasingly smaller sections of society in comparison to the remainder of the classification. The consultation on all names, in particular the Subgroups, was therefore of paramount importance. This can be a time consuming process and with the ONS keen on publishing the 2011 OAC as soon as possible, an alternative method of consultation was required. In advance of the release of the equivalent Scottish Census data in December 2013, the 'Preliminary 2011 England and Wales Area Classification for Output Areas' or 2011 EW OAC, was created prior to the UK-wide 2011 OAC. This classification was constructed using the same methods as would be used for the UK-wide classification, including the naming of the 8 Supergroups,

24 Groups and 67 Subgroups (see Table C.2 in Appendix C). The names assigned here could subsequently be applied to the UK classification and assessed for their suitability.

The 2011 EW OAC was made publically available in September 2013 via the www.retailresearchdata.org website. This allowed the data to be downloaded and the cluster assignments to be visualised using an interactive map. Users who accessed this resource were asked to provide feedback on the classification, most importantly on the naming structure. This informal feedback exercise was independent from the 2011 OAC user engagement and ran over a three-week period. In total there were 11 respondents, six of which answered the question relating to the naming conventions used. Of these responses, two identified potential issues associated with the incorrect interpretation and incorrect inference from the assigned names. Respondent 3 noted that some of the names chosen went beyond the information provided by the 2011 UK Census data. An example they gave was the application of the name 'Industrial Legacy' to areas which have never had any industry. Respondent 10 suggested that the names of the clusters needed to be broader at the Supergroup and Group level, otherwise the Subgroup names started to become too similar. In general, a number of 2011 EW OAC Subgroup names incorporated aspects of their parent Group or Supergroup. However, this was a necessary step in the creation of enough unique names that offered some descriptive power.

As a result of this preliminary feedback, the process of naming the 2011 OAC clusters was modified. It was decided that names would only initially be allocated at the Supergroup and Group level, which would be made available via the www.retailresearchdata.org website. Feedback was requested from users only if they felt that a cluster name was obviously incorrect. This provided an opportunity for mistakes to be rectified prior to the full release and increase the likelihood that Subgroup names would subsequently be accepted by the potential user base. It was only after this feedback had been received that the names for the 76 Subgroups were finalised.

Overall, the creation of the 2011 EW OAC proved invaluable in the creation of names that could be transferred to the 2011 UK wide OAC. It also highlighted the impact that cluster names have on user interpretation of geodemographic classifications. The names for the 2011 OAC Supergroups, Groups and Subgroups are shown in Table 7.10.

Table 7.10: The names for the 2011 OAC Supergroups, Groups and Subgroups

Supergroup	Group	Subgroup
1 - Rural Residents	1a - Farming Communities	1a1 - Rural Workers and Families
		1a2 - Established Farming Communities
		1a3 - Agricultural Communities
		1a4 - Older Farming Communities
	1b - Rural Tenants	1b1 - Rural Life
		1b2 - Rural White-Collar Workers
		1b3 - Ageing Rural Flat Tenants
	1c - Ageing Rural Dwellers	1c1 - Rural Employment and Retirees
		1c2 - Renting Rural Retirement
		1c3 - Detached Rural Retirement
2 - Cosmopolitans	2a - Students Around Campus	2a1 - Student Communal Living
		2a2 - Student Digs
		2a3 - Students and Professionals
	2b - Inner-City Students	2b1 - Students and Commuters
		2b2 - Multicultural Student Neighbourhoods
	2c - Comfortable Cosmopolitans	2c1 - Migrant Families
		2c2 - Migrant Commuters
		2c3 - Professional Service Cosmopolitans
	2d - Aspiring and Affluent	2d1 - Urban Cultural Mix
		2d2 - EU White-Collar Workers
		2d3 - Highly-Qualified Quaternary Workers
3 - Ethnicity Central	3a - Ethnic Family Life	3a1 - Established Renting Families
		3a2 - Young Families and Students
	3b - Endeavouring Ethnic Mix	3b1 - Striving Service Workers
		3b2 - Bangladeshi Mixed Employment
		3b3 - Multi-Ethnic Professional Service Workers
	3c - Ethnic Dynamics	3c1 - Constrained Neighbourhoods
		3c2 - Constrained Commuters
	3d - Aspirational Techies	3d1 - Established Tech Workers
		3d2 - Old EU Tech Workers
		3d3 - New EU Tech Workers
4 - Multicultural Metropolitans	4a - Rented Family Living	4a1 - Social Renting Young Families
		4a2 - Private Renting New Arrivals
		4a3 - Commuters with Young Families
	4b - Challenged Asian Terraces	4b1 - Asian Terraces and Flats
		4b2 - Pakistani Communities
	4c - Asian Traits	4c1 - Achieving Minorities
		4c2 - Multicultural New Arrivals
		4c3 - Inner City Ethnic Mix
5 - Urbanites	5a - Urban Professionals and Families	5a1 - White Professionals
		5a2 - Multi-Ethnic Professionals with Families
		5a3 - Families in Terraces and Flats
	5b - Ageing Urban Living	5b1 - Delayed Retirement
		5b2 - Communal Retirement
		5b3 - Self-Sufficient Retirement

Supergroup	Group	Subgroup
6 - Suburbanites	6a - Suburban Achievers	6a1 - Indian Tech Achievers
		6a2 - Comfortable Suburbia
		6a3 - Detached Retirement Living
		6a4 - Ageing in Suburbia
	6b - Semi-Detached Suburbia	6b1 - Multi-Ethnic Suburbia
		6b2 - White Suburban Communities
		6b3 - Semi-Detached Ageing
		6b4 - Older Workers and Retirement
7 - Constrained City Dwellers	7a - Challenged Diversity	7a1 - Transitional Eastern European Neighbourhoods
		7a2 - Hampered Aspiration
		7a3 - Multi-Ethnic Hardship
	7b - Constrained Flat Dwellers	7b1 - Eastern European Communities
		7b2 - Deprived Neighbourhoods
		7b3 - Endeavouring Flat Dwellers
	7c - White Communities	7c1 - Challenged Transitionaries
		7c2 - Constrained Young Families
		7c3 - Outer City Hardship
	7d - Ageing City Dwellers	7d1 - Ageing Communities and Families
		7d2 - Retired Independent City Dwellers
		7d3 - Retired Communal City Dwellers
		7d4 - Retired City Hardship
8 - Hard-Pressed Living	8a - Industrious Communities	8a1 - Industrious Transitions
		8a2 - Industrious Hardship
	8b - Challenged Terraced Workers	8b1 - Deprived Blue-Collar Terraces
		8b2 - Hard-Pressed Rented Terraces
	8c - Hard-Pressed Ageing Workers	8c1 - Ageing Industrious Workers
		8c2 - Ageing Rural Industry Workers
		8c3 - Renting Hard-Pressed Workers
	8d - Migration and Churn	8d1 - Young Hard-Pressed Families
		8d2 - Hard-Pressed Ethnic Mix
		8d3 - Hard-Pressed European Settlers

7.4.3. Cluster descriptions

The naming of clusters is important as it provides an easily understood 'snapshot' of each group. However, the names do not allow for a comprehensive understanding of the resident population and physical characteristics in areas. Cluster descriptions, or pen portraits provide this more detailed information, and are therefore an integral part of any geodemographic classification. Similar to the naming of clusters, pen portraits were easier to construct for clusters demonstrating more extreme values. In these instances pen portraits are particularly relevant as the dynamics within a cluster would be too great to encapsulate in a name alone.

Where clusters lacked extreme values, it was necessary to consider their geographic locations. Although two clusters may represent 'average' characteristics, if their spatial distribution varies, their individual characteristics will vary too. The pen portraits provided an opportunity to explore the dynamics of these clusters further. The pen portraits for the 2011 OAC are included in Section C.1 in Appendix C.

The nesting of Subgroups within different Groups, and Groups within Supergroups, in the 2011 OAC means there are two different ways of interpreting the average characteristics of these clusters. Comparisons between pen portraits are therefore not straightforward and require an understanding of what data is being used to draw conclusions about the characteristics of each cluster.

These different methods of interpretation are a result of there being two possible mean values for each Group and Subgroup. The first possible mean value refers to only that clusters parent Supergroup or Group. These mean values are impacted by the prevalence of the particular variable in their respective parent cluster. For example, two Groups from different parent Supergroups may demonstrate identical values of variable *x*. However, one of the parent Supergroups may actually contain below the national average of variable *x*, whilst the Group demonstrates above the national average of the variable. The other Group and Supergroup may demonstrate the opposite, with a prevalence of variable *x* above the national average, whilst the Group demonstrates below the national average. In this example both relationships lead to the same mean value for the Groups.

The second possible mean value, like those calculated for the Supergroups, is in relation to the UK average. Using the national average for all the Groups and Subgroups means all 110 clusters of the 2011 OAC can be directly compared with each other. The disadvantage of using this method is it makes it harder to distinguish between Groups and Subgroups that are derived from the same Supergroup or Group.

The two different methods of interpreting the average characteristics of the 2011 OAC clusters can be used in conjunction with each other to provide a detailed understanding of each group. Although there are limitations to both methods, any issues that exist are outweighed by the benefits derived from being able to construct detailed descriptions of the clusters. If the pen portraits for the 2011 OAC are however deemed unsatisfactory for whatever reason by the user, the open methodology allows for alternatives to be

created. All of the materials used to create the pen portraits will also be made available, which provides users with the opportunity to develop their own pen portraits if required.

7.4.4. Mapping the 2011 OAC

The mapping of the 2011 OAC is the second core output of the classification. It provides a visual interface for the cluster names and descriptions and allows users to explore the spatial distribution of the cluster assignments. Documented in this section are maps of the 2011 OAC Supergroups, Groups and Subgroups, although the Groups and Subgroups are better suited to visualisation on an online interactive map.

An important consideration in the production of the output maps was the colour schemes (Harrower and Brewer, 2003; Gardner, 2005). All of the maps presented in this section are based on schemes from www.colorbrewer2.org (see Figure 7.22). This website offers advice on colour schemes to aid the selection of good colour schemes for maps and other graphics based on Brewer (1994). The selected scheme was chosen to be as colour blind safe as possible when visualising eight distinctive clusters, and assigned a colour to each Supergroup. The Group and Subgroup colours were subsequently based on their parent Supergroup, with each cluster being made either lighter or darker to create 26 and 76 different colours respectively. The adaption of 26 and 76 unique colours meant that all the Groups and Subgroups could have been visualised on a two maps of the UK. However, this would be too difficult to interpret, and as a result 16 separate maps have been produced, one for every Group and Subgroup belonging to each of the parent Supergroups.

In total three different ways of mapping the data are presented in this section: choropleth, density equalising cartogram and building maps. A choropleth map provides an accurate representation of the desired areal unit geography, OAs and SAs for the 2011 OAC. Each areal unit displays an assigned category, such as their assigned Supergroup or Group. In contrast, a density equalising cartogram map modifies the size and shape of the areal units based on a specific set of criteria, whilst maintaining geographical accuracy as far as is practical. Finally, a building map takes advantage of the release of Ordnance Survey Open Data (OS OpenData) under the UK Open Government Licence (OGL). The building layer of the 'OS VectorMap District' dataset can be spatially joined to OAs in Great Britain. This assigns every building within Great Britain to either the OA it

falls within, or the OA that contains the largest proportion of the building, if it straddles multiple OAs. Instead of colouring a whole areal unit in the colour of its assigned Supergroup or Group, only the buildings are coloured.

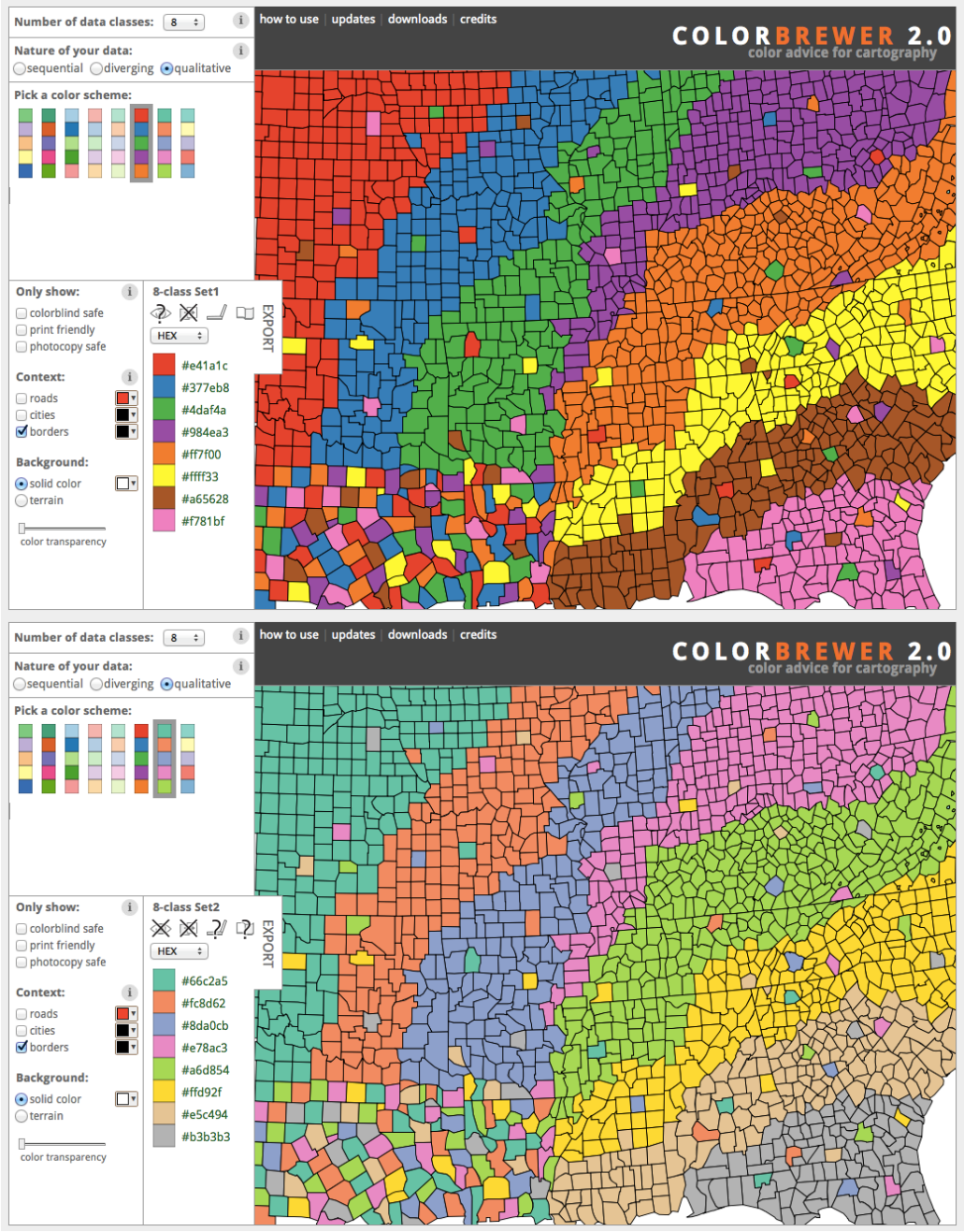


Figure 7.22: Screenshots of www.colorbrewer2.org

7.4.4.1. Choropleth maps

Figure 7.23 shows the distribution of 2011 OAC Supergroups across the UK. Visually, the 'Rural Residents' Supergroup dominates the map, even though it is only assigned to 12% of the UK's OAs and SAs. Spatial clustering of the 'Cosmopolitans', 'Ethnicity Central' and 'Multicultural Metropolitans' can be identified around larger urban areas such as London, Birmingham and Manchester, although they are also present in some of the smaller cities in England. The 'Hard-Pressed Living' Supergroup is most keenly clustered around the valleys of South Wales, and in the North East of England. It is also found in less concentrated areas across other parts of England, such as the Midlands. The 'Constrained City Dwellers' Supergroup is spatially clustered in similar locations to the 'Hard-Pressed Living' Supergroup, except it is more likely to be found in inner urban areas of which Glasgow has the highest concentration.

The 'Suburbanites' Supergroup has a strong spatial clustering around London, especially to the South West. It is also found in lower concentrations on the outskirts of other urban centres, and in on the Isle of Lewis in Scotland. Finally, the only discernible spatial pattern of the 'Urbanities' Supergroup is that it is predominantly located in urban areas across the UK. The middle layer of the 2011 OAC hierarchy is shown in Figures 7.24 to 7.31, and the bottom layer of the hierarchy is shown in Figures 7.32 to 7.39. The distribution of the 26 Groups and 76 Subgroups across the UK provides an opportunity for the spatial patterns seen with the Supergroups to be explored further. The geographic variations between the Groups and Subgroups in particular become more evident when viewed at the sub-UK level.

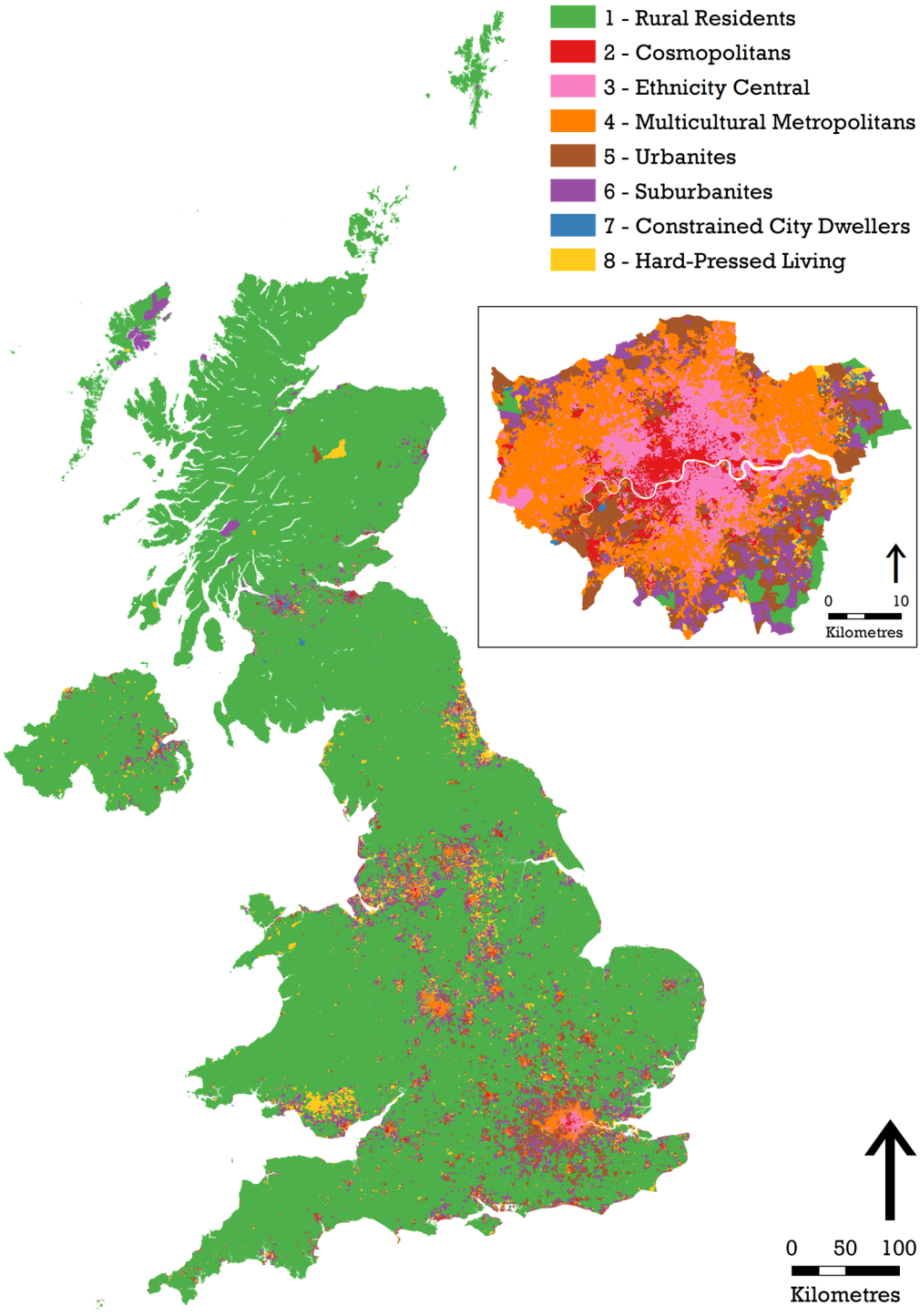


Figure 7.23: Choropleth map of the 2011 OAC Supergroups

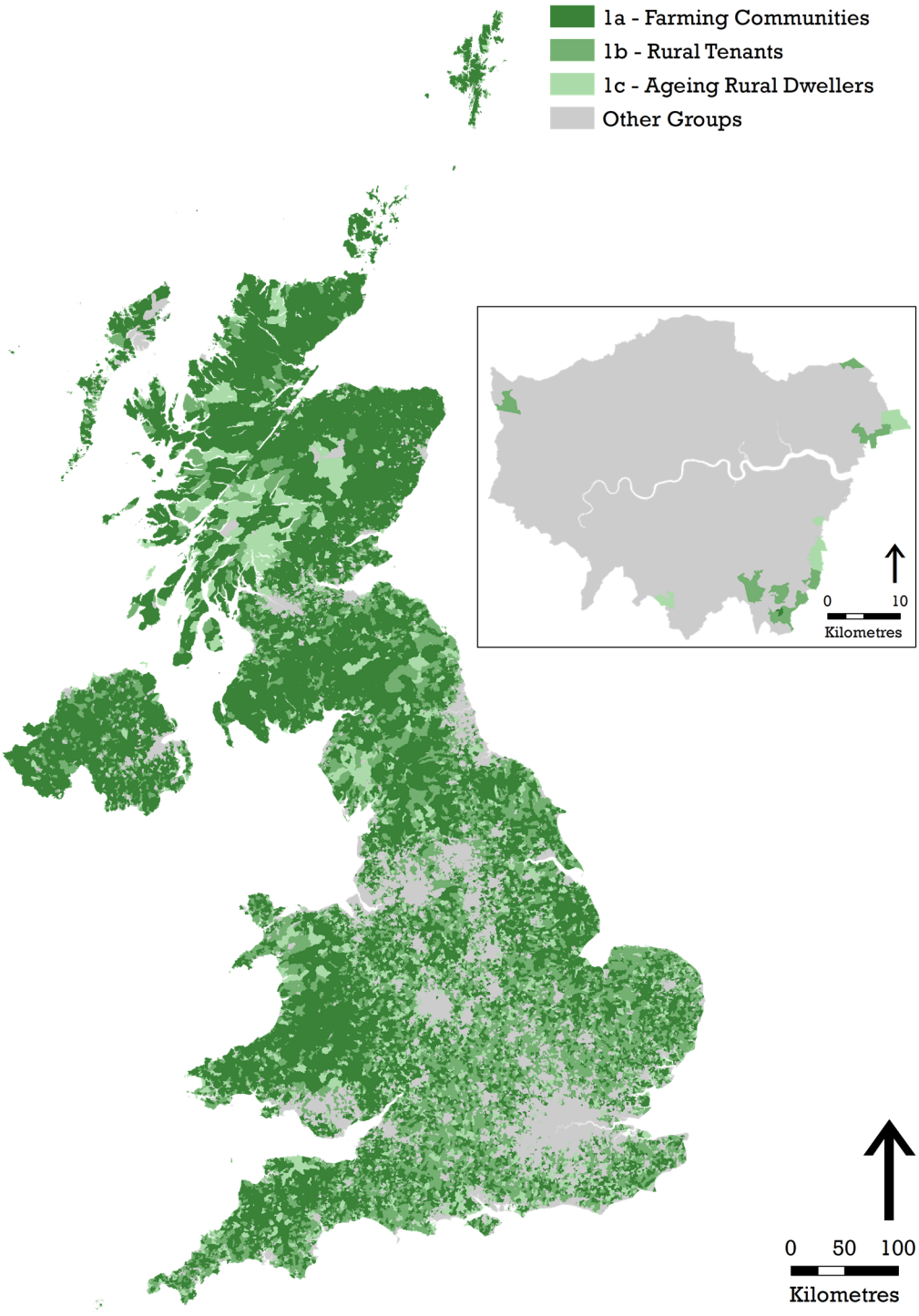


Figure 7.24: Choropleth map of the 2011 OAC Groups derived from the 'Rural Residents' Supergroup

Figure 7.25: Choropleth map of the 2011 OAC Groups derived from the 'Cosmopolitans' Supergroup

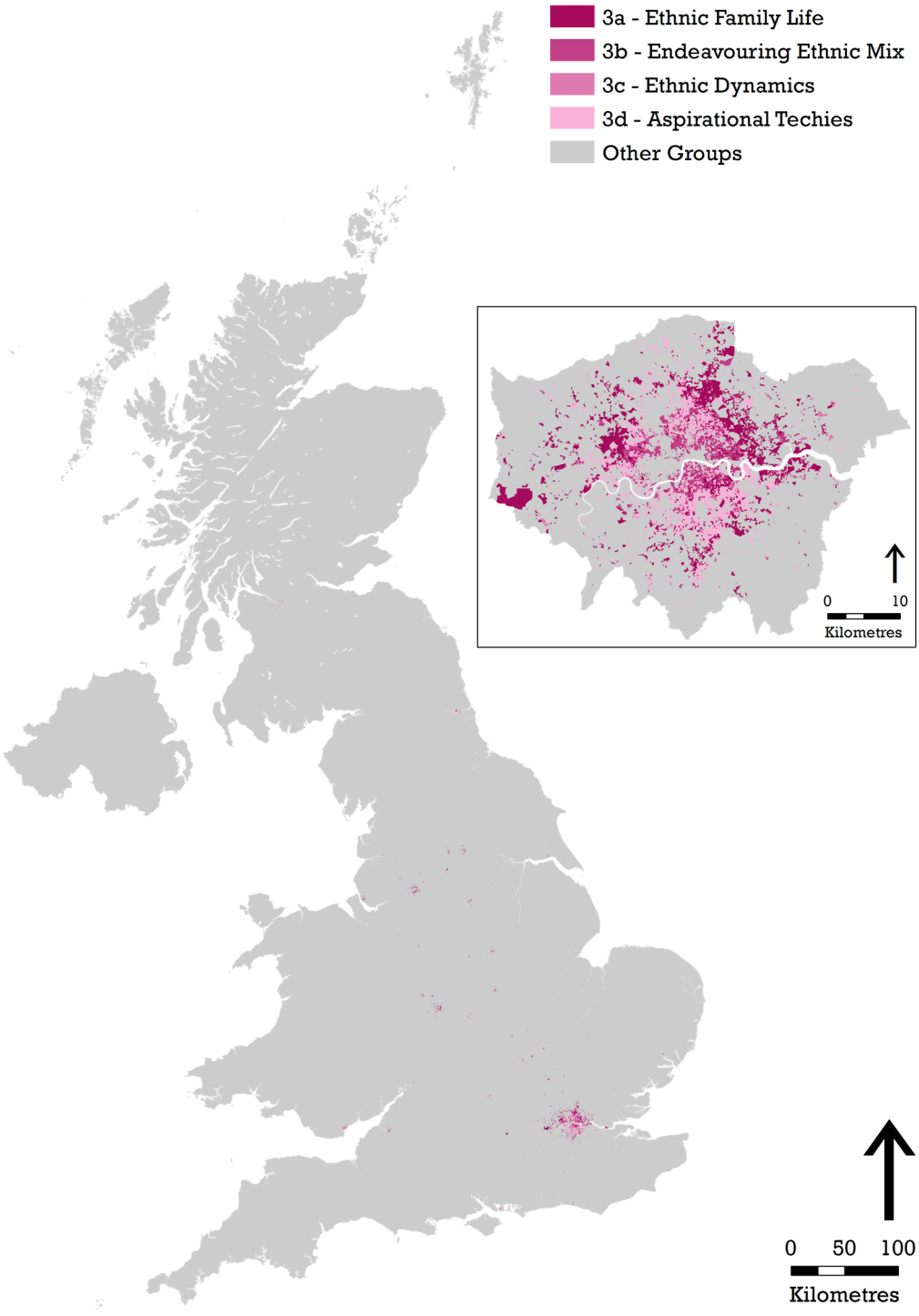


Figure 7.26: Choropleth map of the 2011 OAC Groups derived from the 'Ethnicity Central' Supergroup

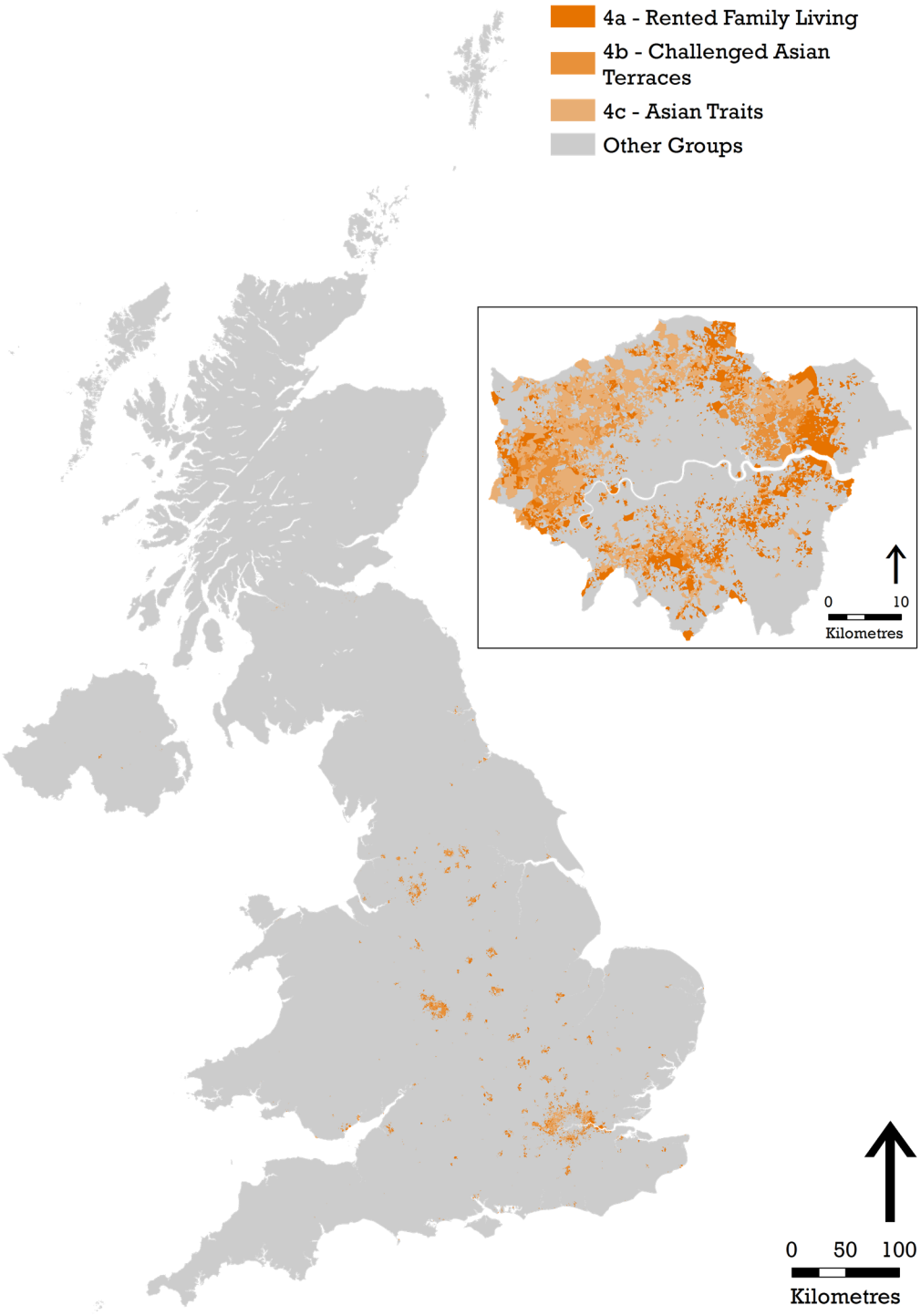


Figure 7.27: Choropleth map of the 2011 OAC Groups derived from the 'Multicultural Metropolitans' Supergroup

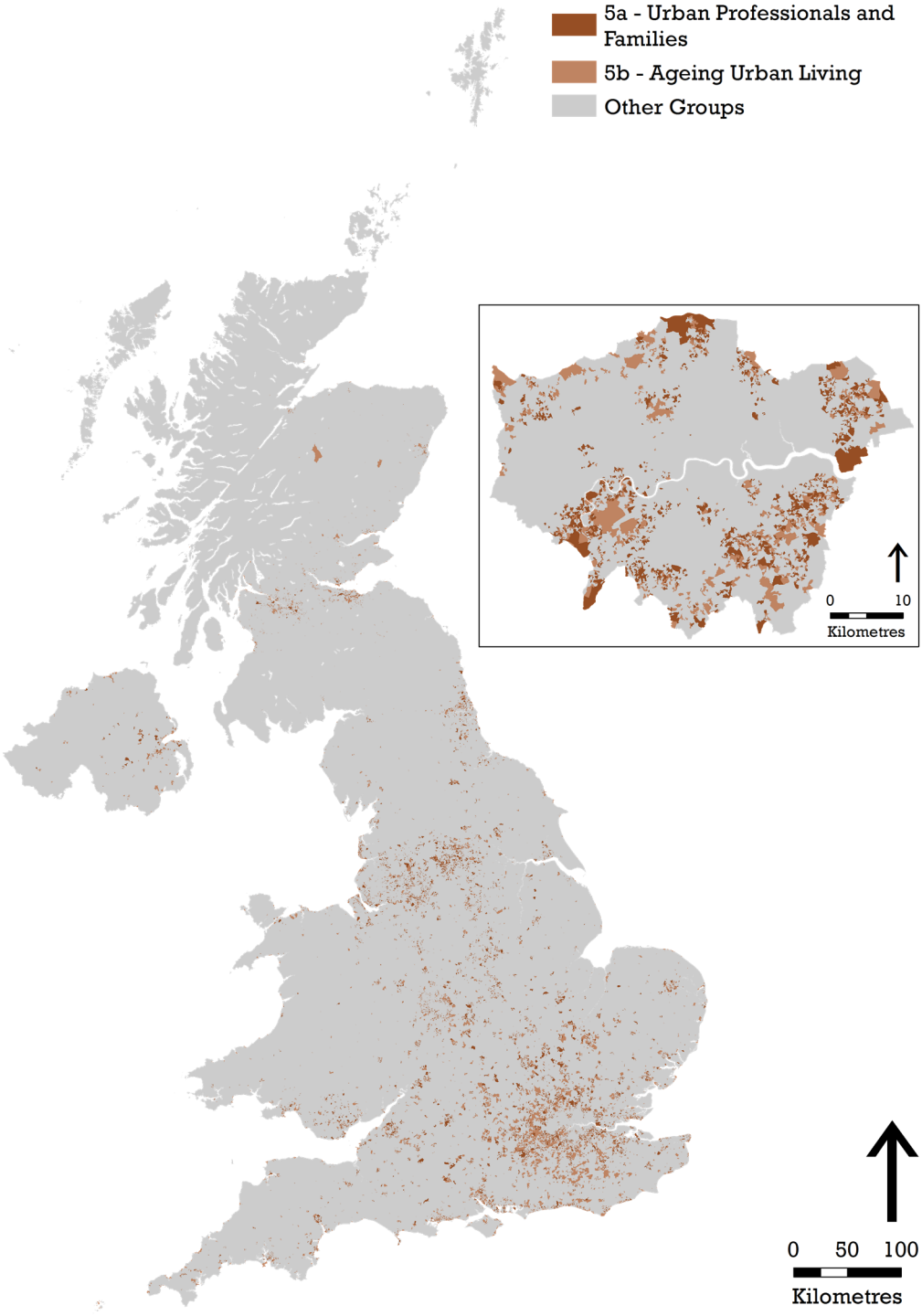


Figure 7.28: Choropleth map of the 2011 OAC Groups derived from the 'Urbanites' Supergroup

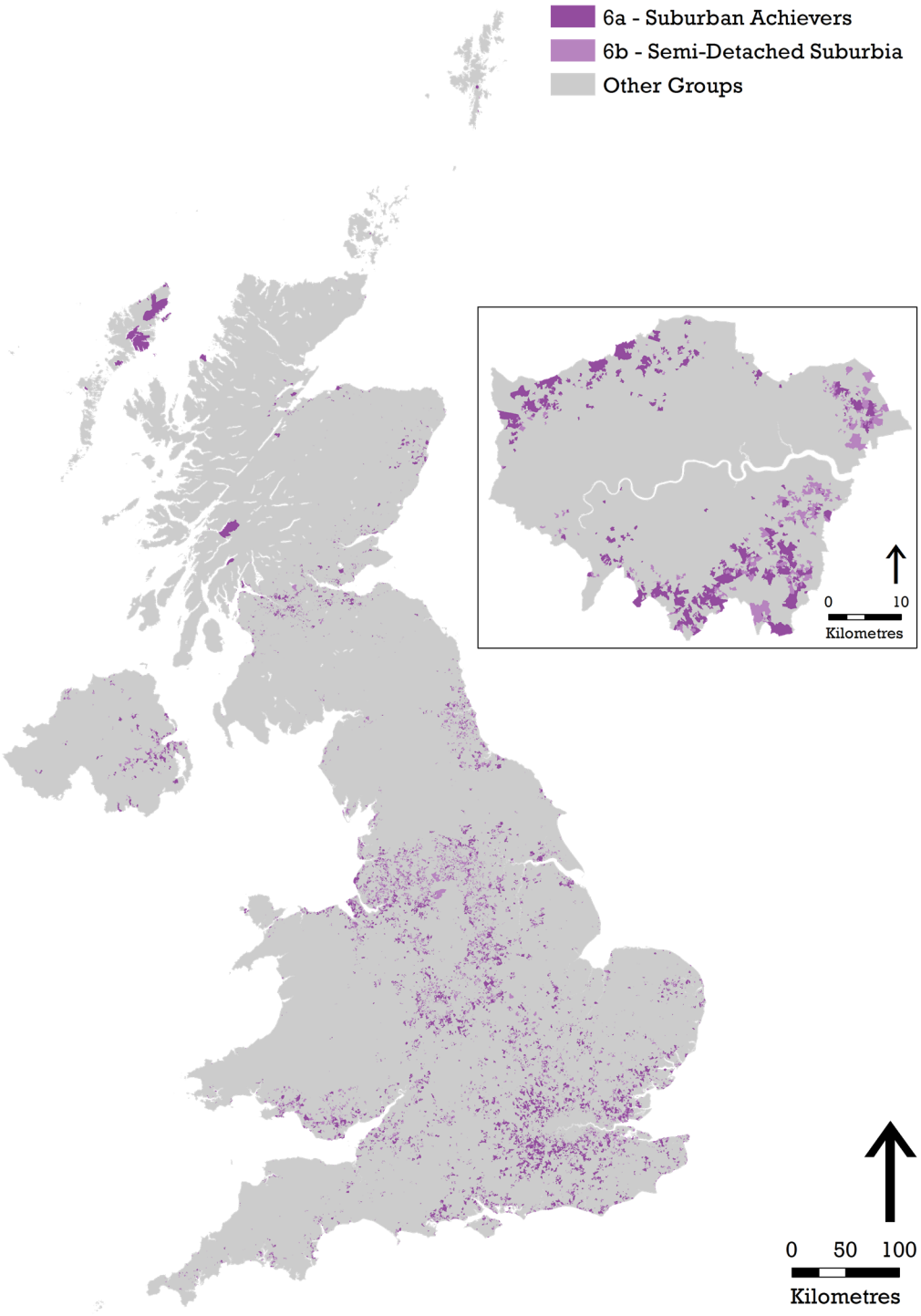


Figure 7.29: Choropleth map of the 2011 OAC Groups derived from the 'Suburbanites' Supergroup

Figure 7.30: Choropleth map of the 2011 OAC Groups derived from the 'Constrained City Dwellers' Supergroup

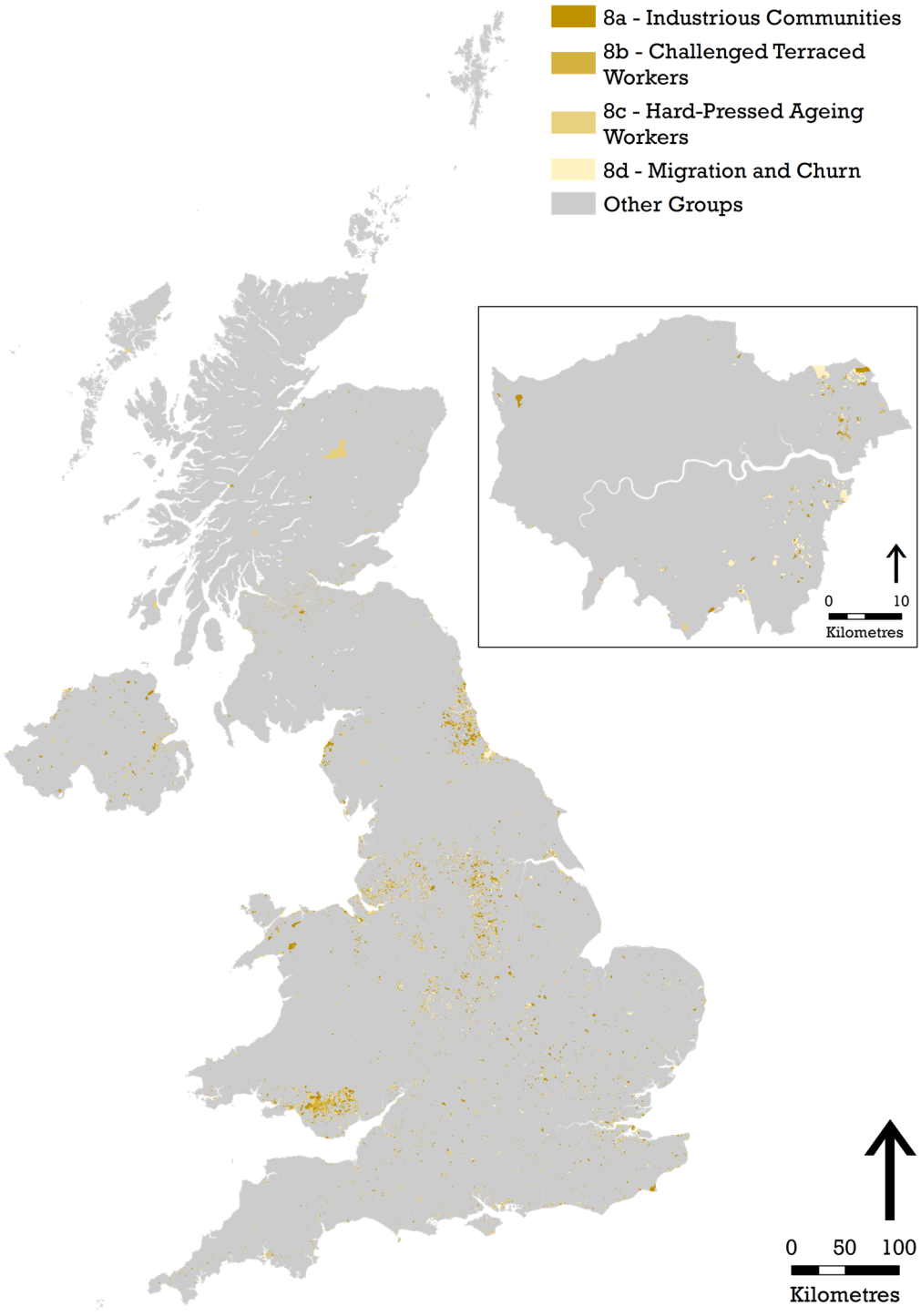


Figure 7.31: Choropleth map of the 2011 OAC Groups derived from the 'Hard-Pressed Living' Supergroup

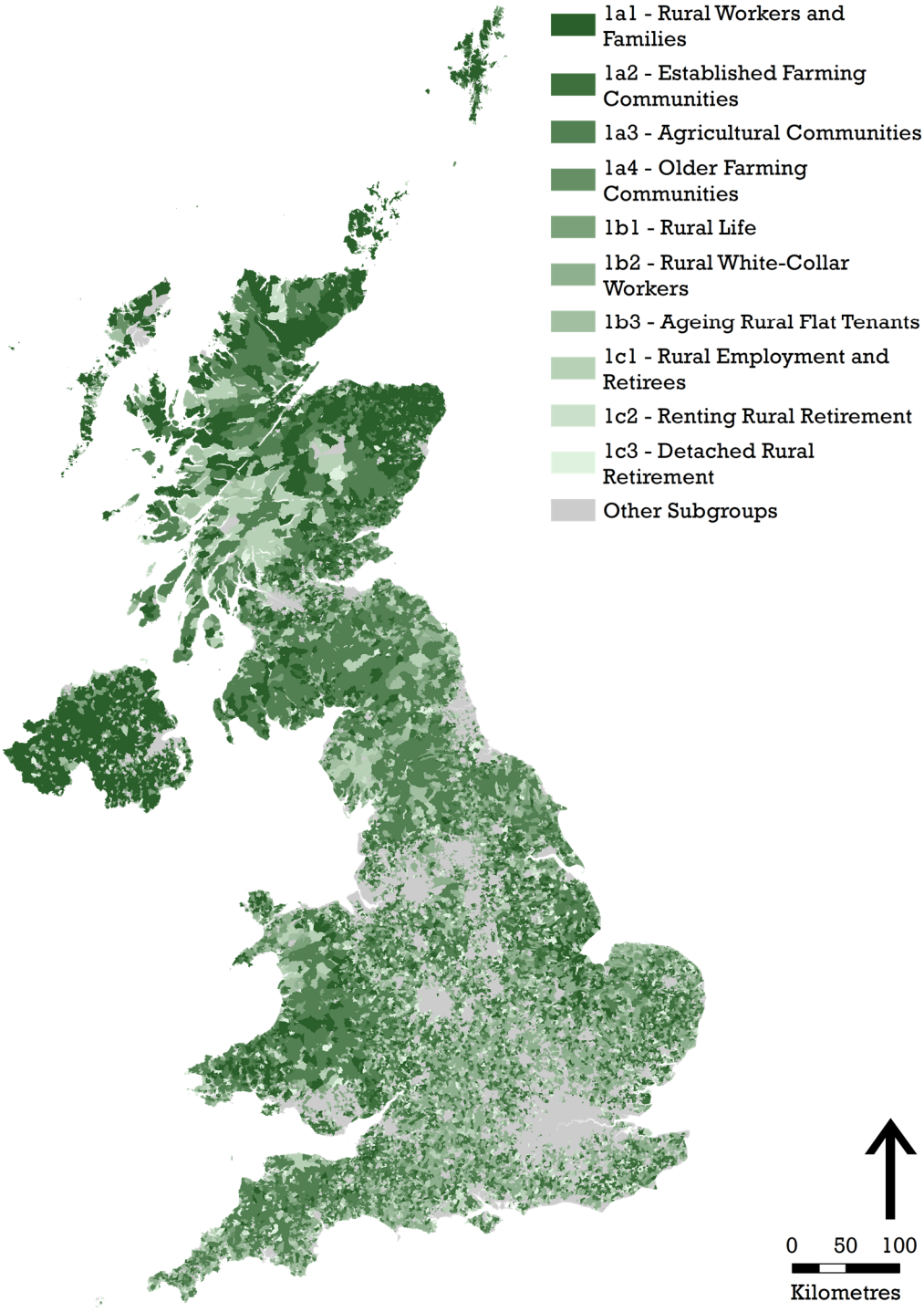


Figure 7.32: Choropleth map of the 2011 OAC Subgroups derived from the 'Rural Residents' Supergroup

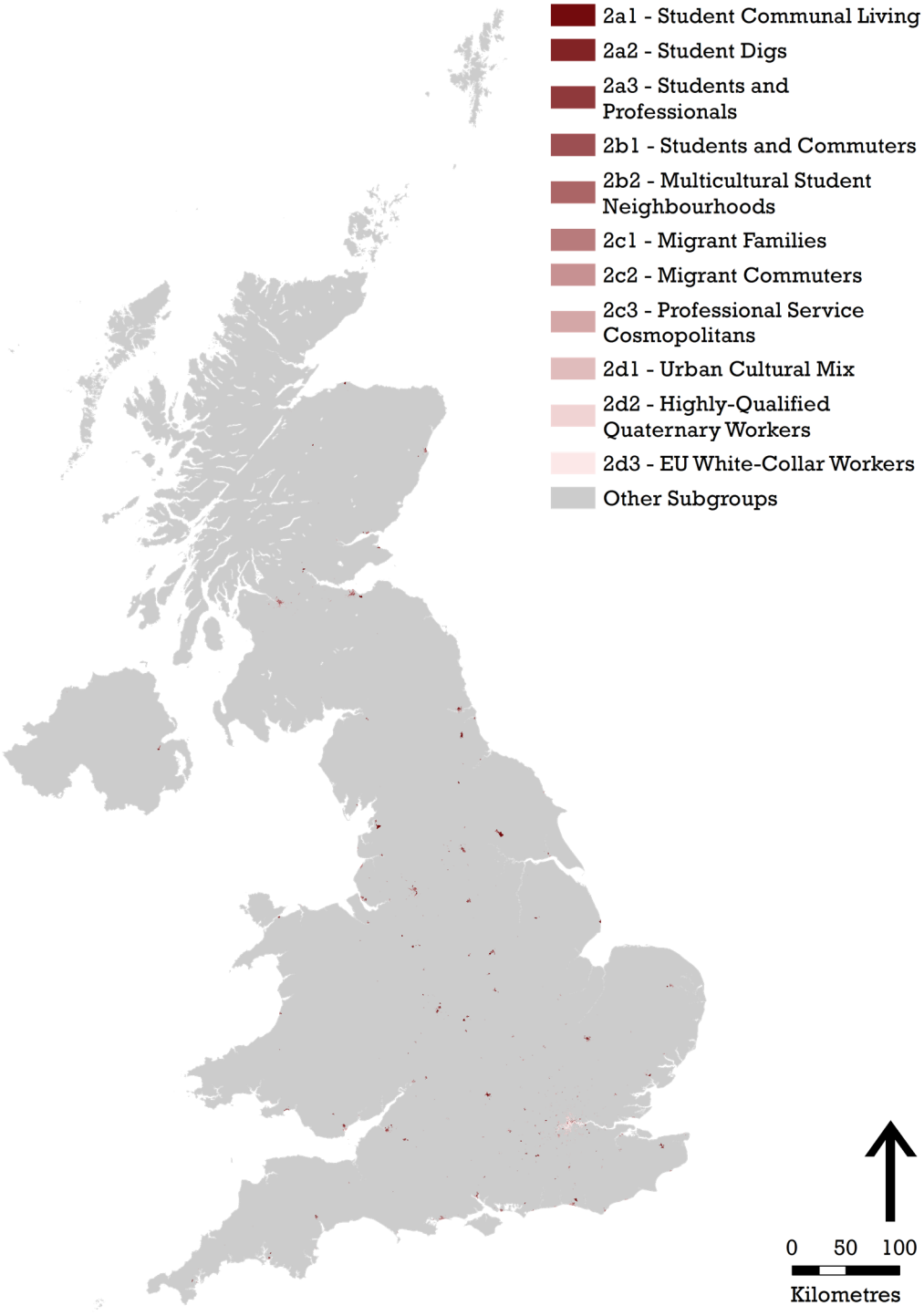


Figure 7.33: Choropleth map of the 2011 OAC Subgroups derived from the 'Cosmopolitans' Supergroup

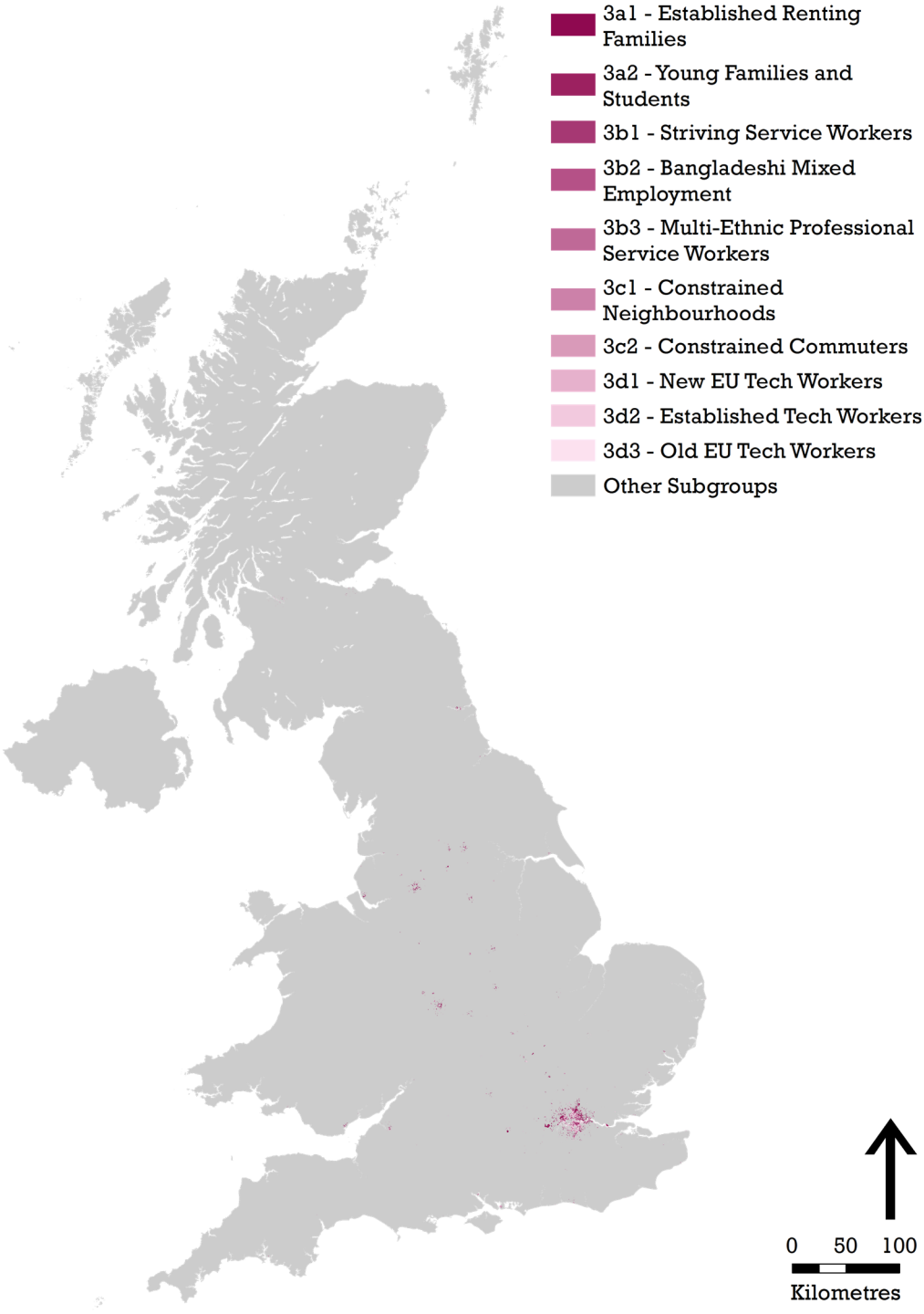


Figure 7.34: Choropleth map of the 2011 OAC Subgroups derived from the 'Ethnicity Central' Supergroup

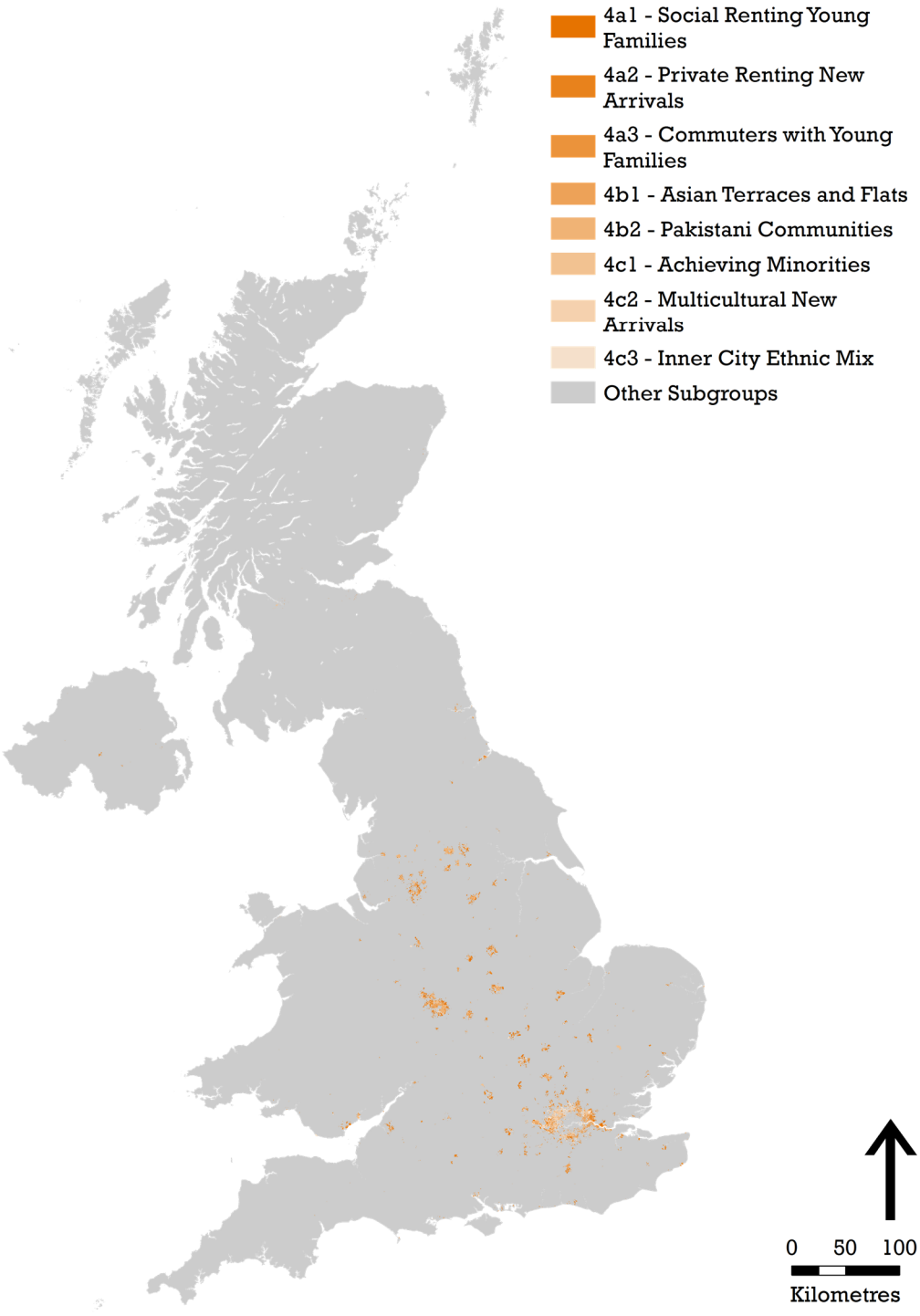


Figure 7.35: Choropleth map of the 2011 OAC Subgroups derived from the 'Multicultural Metropolitans' Supergroup

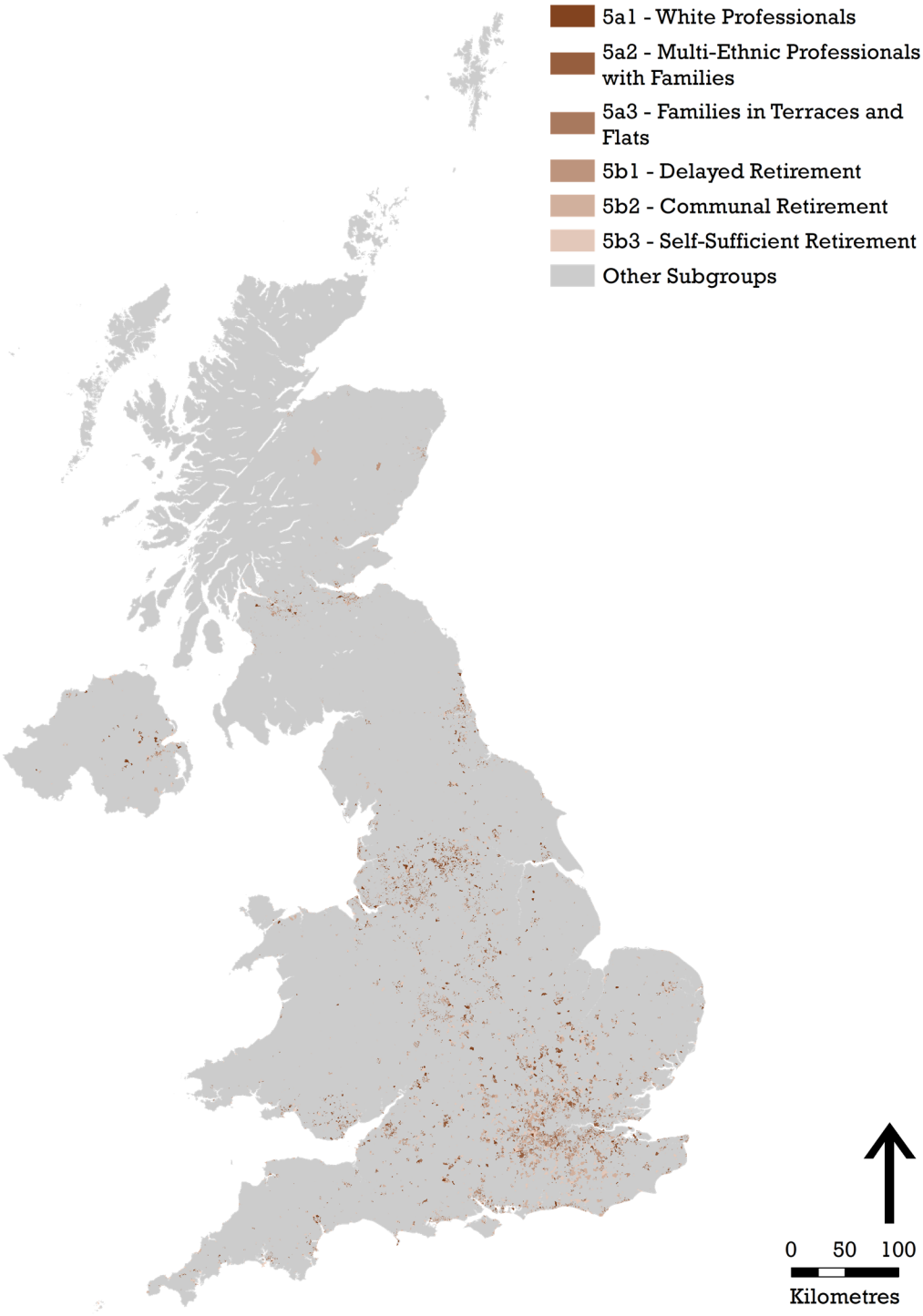


Figure 7.36: Choropleth map of the 2011 OAC Subgroups derived from the 'Urbanites' Supergroup

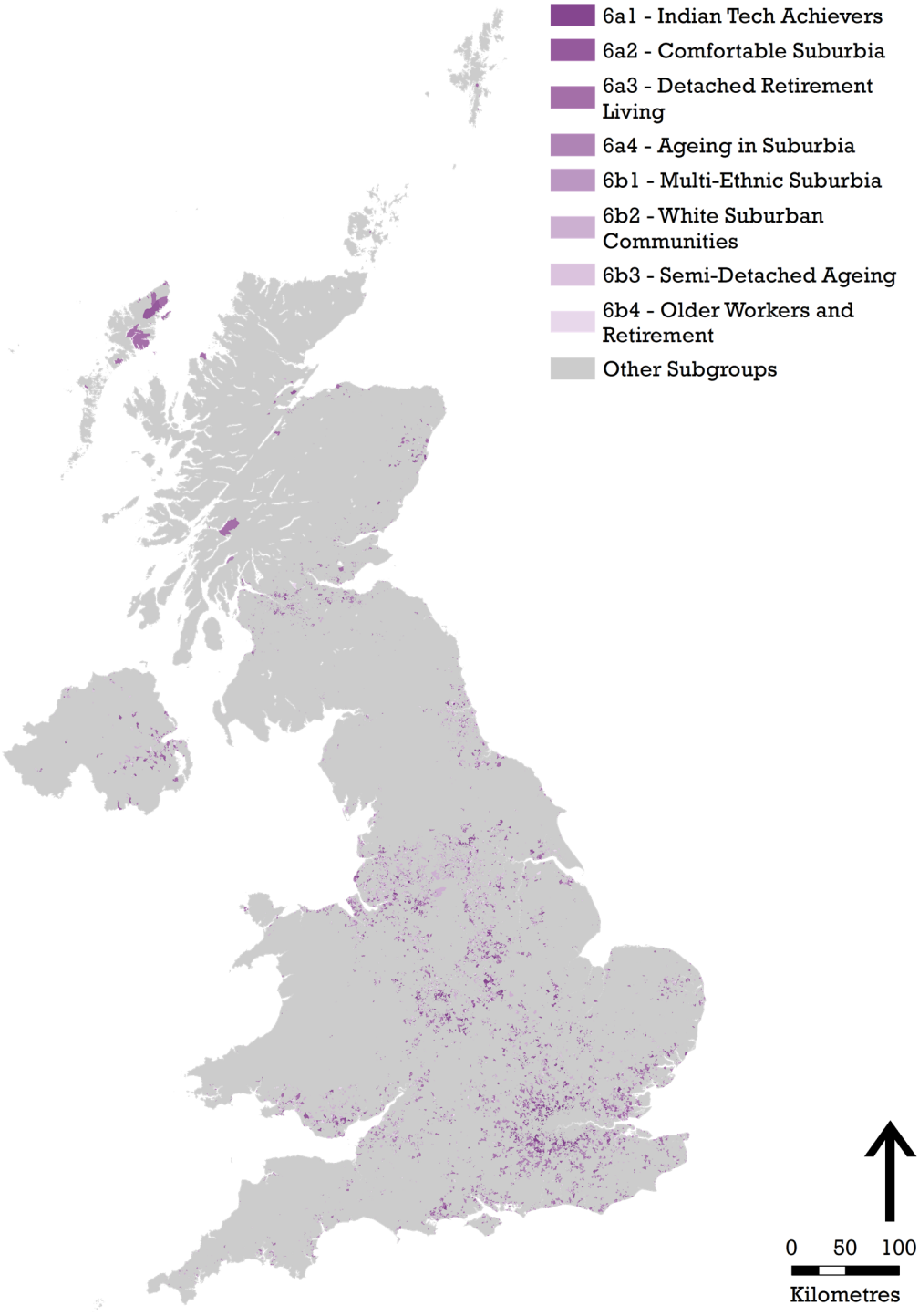


Figure 7.37: Choropleth map of the 2011 OAC Subgroups derived from the 'Suburbanites' Supergroup

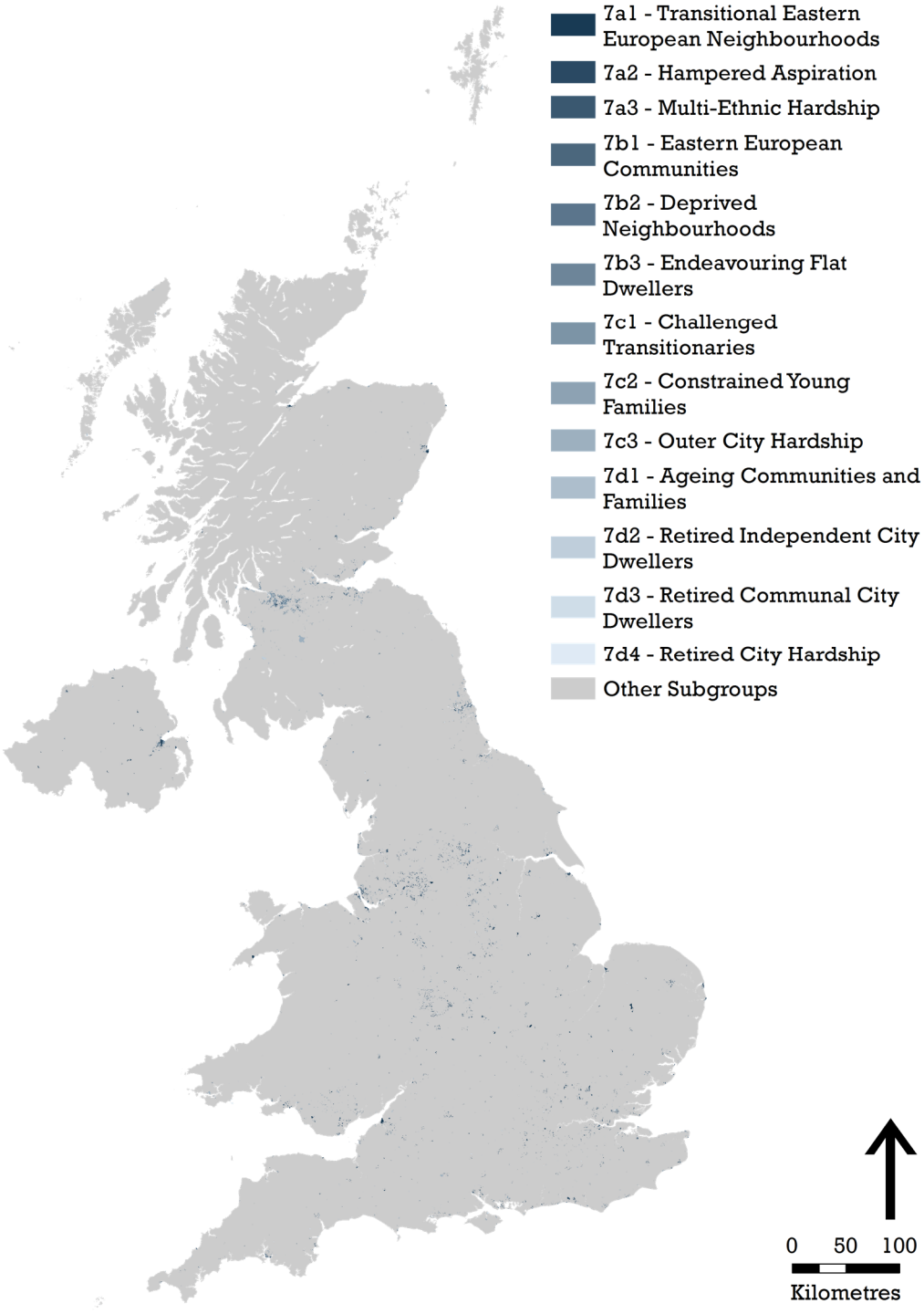


Figure 7.38: Choropleth map of the 2011 OAC Subgroups derived from the 'Constrained City Dwellers' Supergroup

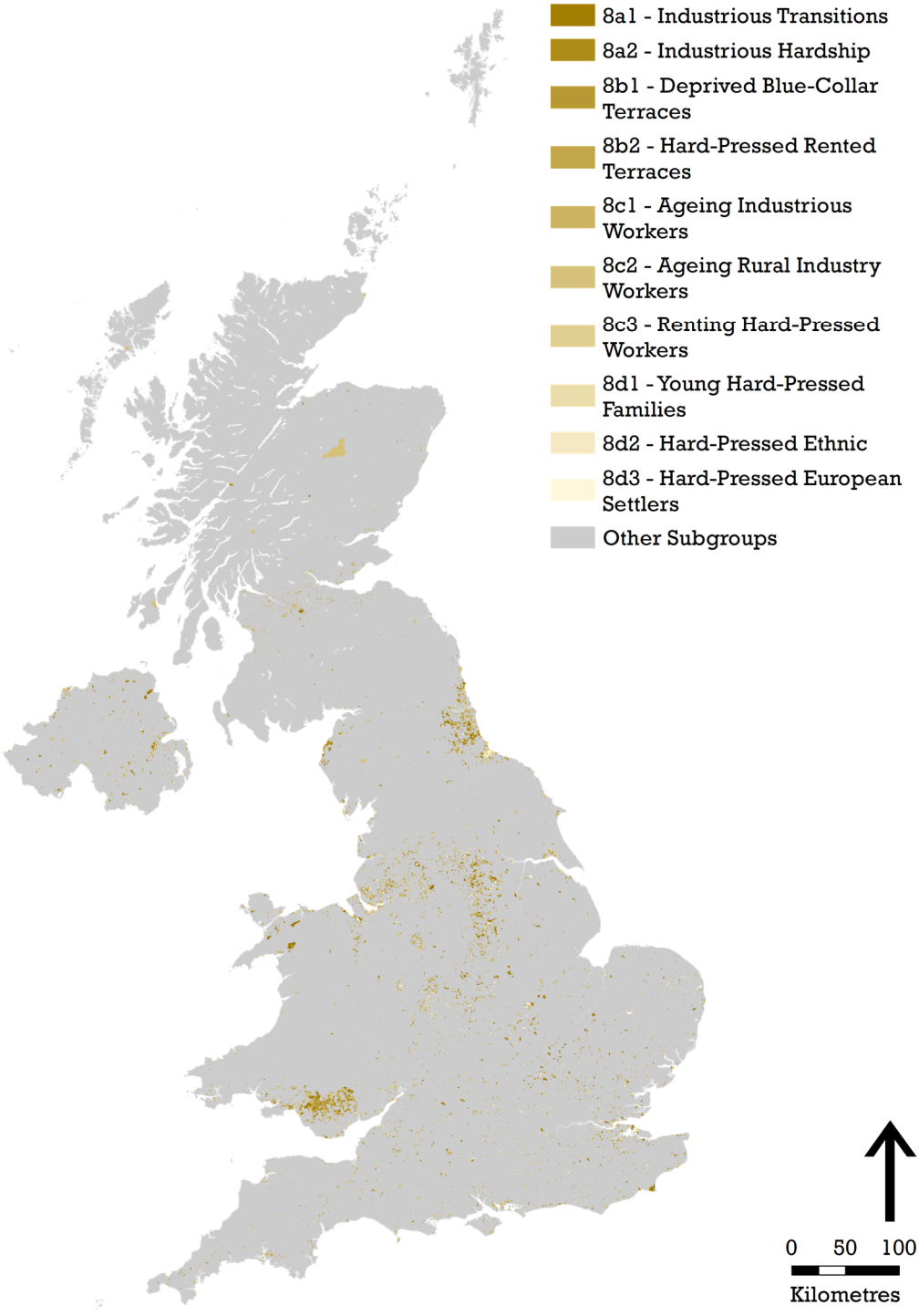


Figure 7.39: Choropleth map of the 2011 OAC Subgroups derived from the 'Hard-Pressed Living ' Supergroup

7.4.4.2. Cartogram maps

The choropleth maps shown in Figures 7.23 to 7.39 provide a standard geographical representation of the UK, and the OAs and SAs they contain. A drawback of this type of visualisation is that it does not distinguish between more densely populated urban areas and less densely population rural areas. This results in a visual domination of the map by the 'Rural Residents' 2011 OAC Supergroup because it happens to represent predominantly rural areas. In total it covers 87% of the UK's total land mass, yet represents only 11.6% of the UK's population. This discrepancy is caused by the OA and SA minimum population threshold values as discussed in Section 3.3.1. It was necessary for such a wide area to be incorporated to meet this minimum value. This imbalance was addressed utilising a density equalising cartogram to modify the areal units to reflect their resident population (Gastner and Newman, 2004).

A density equalising cartogram changes the size and shape of an areal unit based on another attribute. In the case of the 2011 OAC this was the total population of each OA and SA as recorded in the 2011 UK Census. Each OA and SA was resized so that those with the highest resident population would become larger, and those with lower populations would become smaller. To guarantee some geographic stability with the final outcome, the neighbours of each OA and SA remained the same after the process was complete, with no new neighbours or gaps being added. The outcome of this process is shown in Figure 7.40. Two aspects from the map are particularly striking: firstly the increase in the relative sizes of the urban areas in comparison to the choropleth maps; and secondly the prominent difference in population sizes between England and the rest of the UK. The dominance of urban areas, and in particular London, make it a lot easier to identify the spatial patterning of the Supergroups found in these areas, even if the geographic integrity of the choropleth maps has been lost.

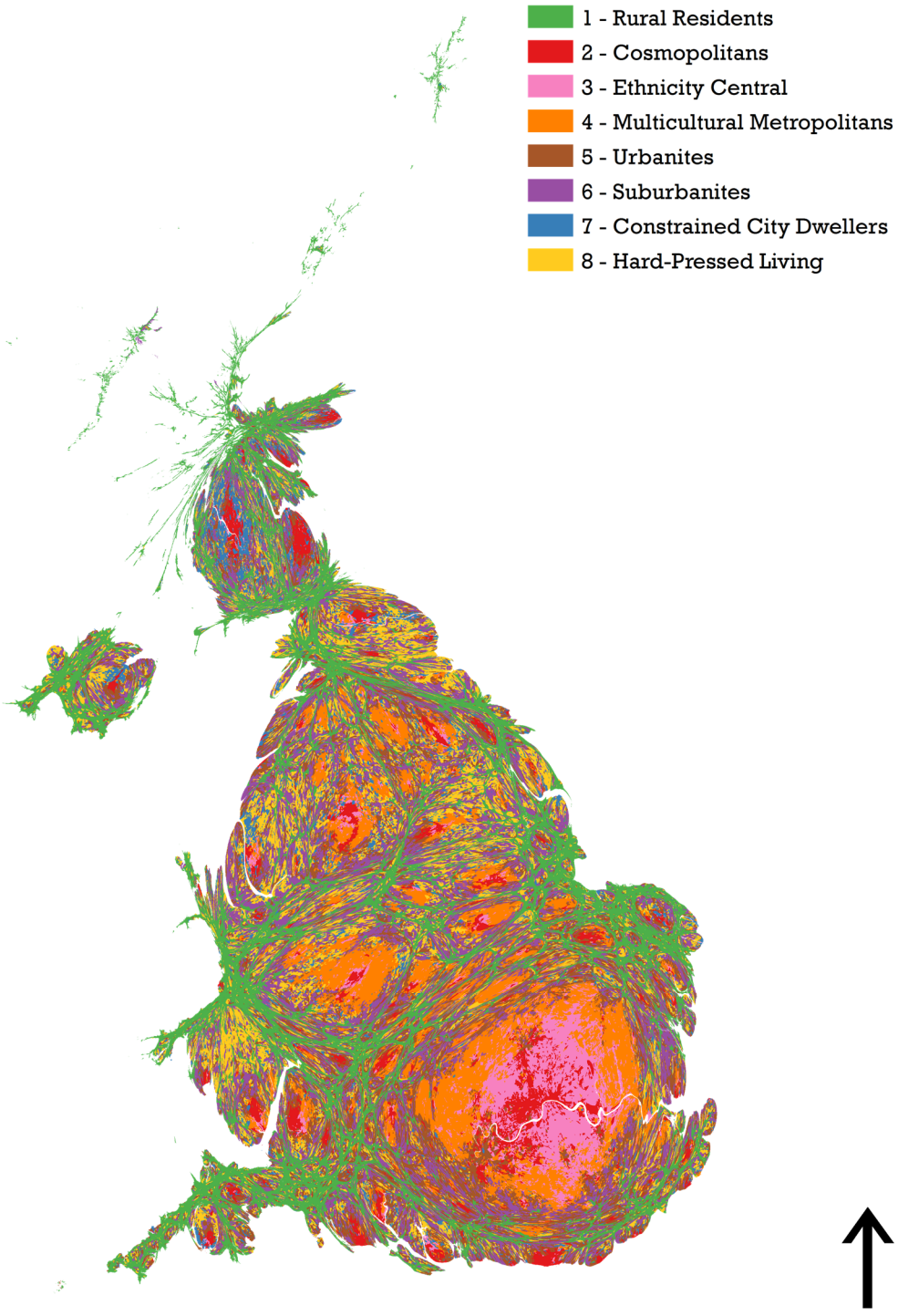


Figure 7.40: Cartogram map of the 2011 OAC Supergroups

Undoubtedly the cartogram technique does re-address the rural/urban divide that exists in the 2011 OAC in regard to population representation. However, in making the urban areas more prominent, the geographic accuracy of the map is reduced. In order to preserve the size and shape of an OA or SA, another is significantly altered. Although the outline of Figure 7.40 resembles the UK, the land mass has become a blur of stretched and squeezed OAs and SAs. Despite the technique retaining an element of geographic stability, no locations, outside of the major urban centres, can be identified. For example, towns and cities on the south coast of England become almost lost due to their relatively small size in comparison to London. The key benefit of the cartograms in comparison to the choropleth maps is that they provide a better visual representation of the UK's eight Supergroups, rather than being dominated by the 'Rural Residents' Supergroup. However, their use is limited when seeking to understand spatial patterns of clustering outside of London.

7.4.4.3. Building maps

The geographic simplicity offered by a choropleth map and the ability of the cartogram to distinguish between urban and rural areas are two positive elements of using these techniques to map the 2011 OAC. However, both types of maps represent an entire OA or SA in the same way irrespective of the number of households that are present within it. The 2011 UK Census dataset, from which the 2011 OAC is constructed, was based on Census returns completed per household. As such, the mapping of locations solely where these households are found is an opportunity to discard other amenities found in OAs or SAs, such as allotments or parks. This was achieved through the incorporation of the building layer from the 'OS VectorMap District' dataset. The assignment of each building to an OA allowed for the visualisation of individual buildings. Figure 7.41 shows a standard choropleth representation of the 2011 OAC Supergroups in London. Figure 7.42 is the same representation but here only the buildings are coloured.

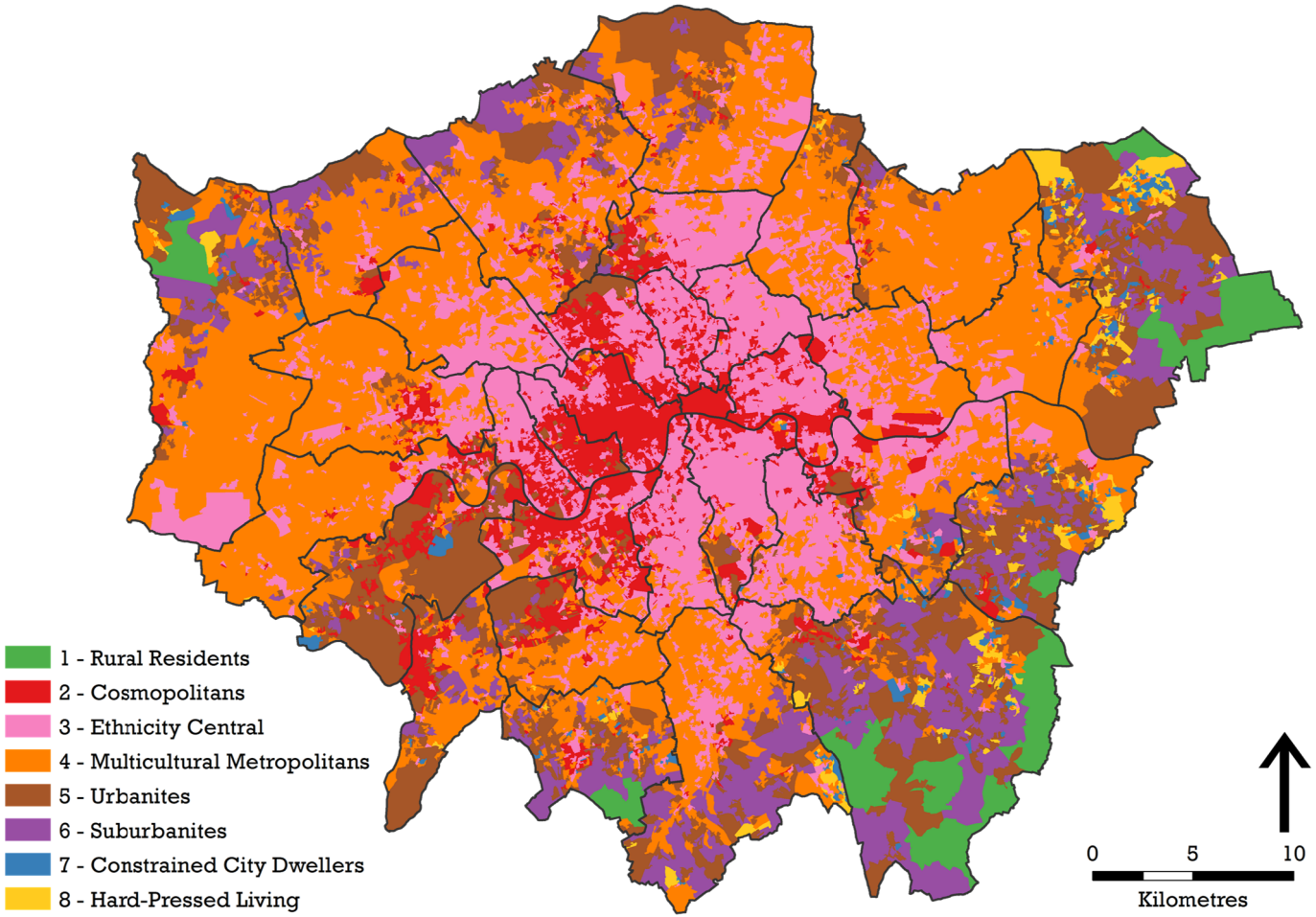


Figure 7.41: Choropleth map of the 2011 OAC Supergroups in London

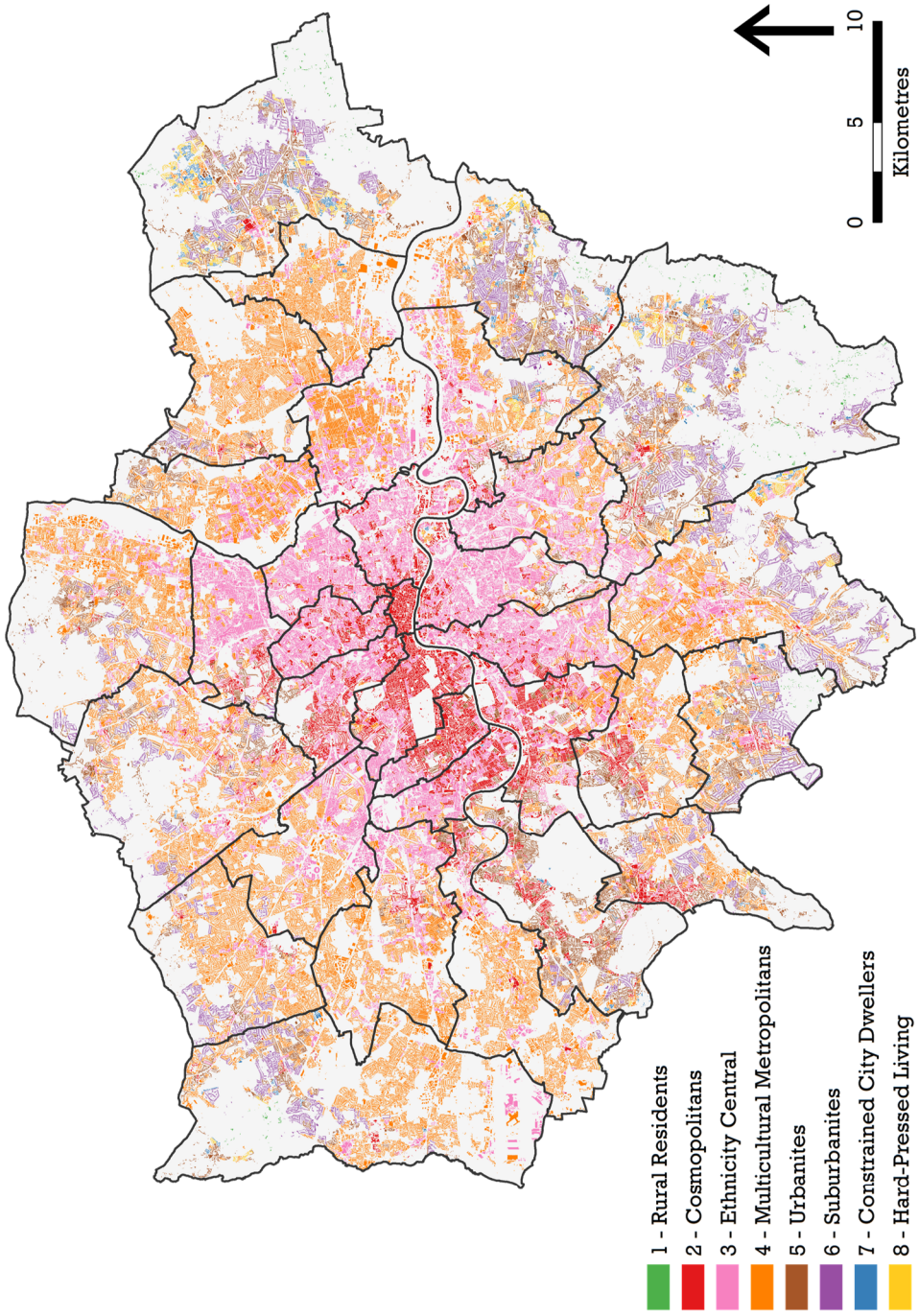


Figure 7.42: Building map of the 2011 OAC Supergroups in London

In contrasting the two maps of London it becomes clear how features hidden by a choropleth map are revealed when only the buildings are represented. A number of London's parks are clearly visible, and so are places like Lea Valley. It also becomes apparent that there are rural areas in the outer London Boroughs. In the choropleth map (Figure 7.41), areas of 'Rural Residents' were few in number but clearly identifiable. In the building map (Figure 7.42), they become almost invisible, revealing how sparsely populated these areas are in comparison to the inner London Boroughs. Figure 7.43 contrasts the choropleth map and building map in central London. At this level individual roads can be identified, along with Buckingham Palace. It also identifies a current flaw with the technique, namely that no distinctions are made between residential and commercial property. Figures 7.42 and 7.43 visualise all buildings indiscriminately, irrespective of their use. This leads to a commercial building being assigned characteristics based on a residential population. The other issue of assigning a single building to a Supergroup, Group or Subgroup is that it can be implied that the classification has been constructed at the building level, rather than OA or SA level. Any attempts to clarify this on the map itself, by adding OA boundaries for example, would create clutter and make the visualisation more complicated to use.

Preventing a user from misinterpreting a map can be a challenging task and this is a particular issue for the building maps of London. They lend themselves to being inappropriately compared with Charles Booth's visualisations of deprivation and poverty (as detailed in Section 2.3) because they are mapping similar entities. This can lead to misinterpretation, as despite the similar appearance, the building maps do not offer the same level of detail as Booth's work, they simply provide an alternative to choropleth and cartogram maps. Geographic integrity is maintained, but the higher population densities in urban areas becomes apparent. At a national level this form of visualisation is better suited to being provided as online interactive map. The lack of buildings (when compared to towns and cities) means that large areas of the UK's countryside are left blank. Although the mapping of buildings is not necessarily the best way to visualise the 2011 OAC for all users, it provides a more useful alternative to a standard choropleth map than that offered by a cartogram. It is however likely that each visualisation method discussed will be favoured by different users of the 2011 OAC, with this dependent on their intended use for the classification.

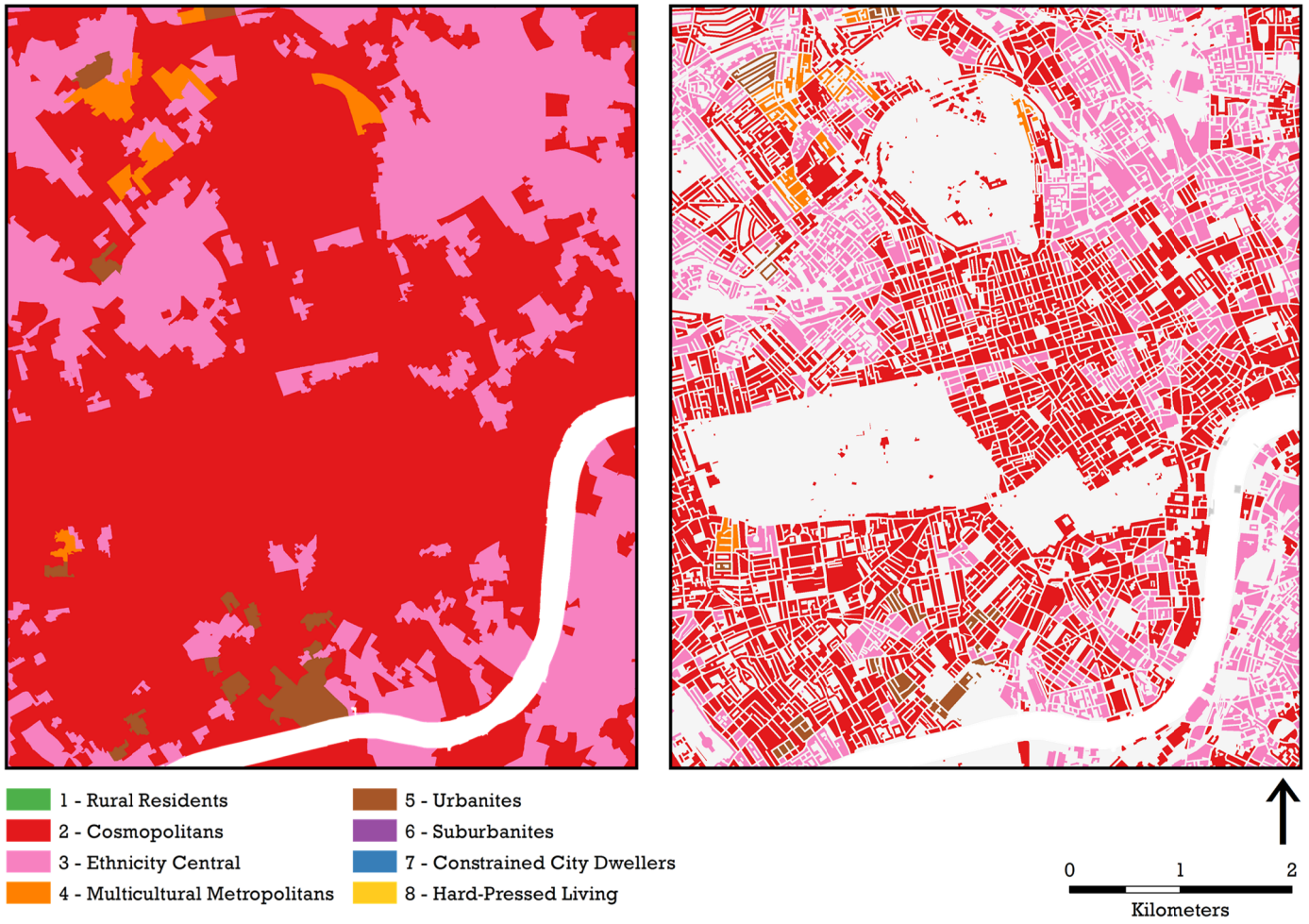


Figure 7.43: Choropleth and building comparison maps of the 2011 OAC Supergroups in central London

7.4.5. Other Outputs

In addition to the descriptive and visual outputs of the 2011 OAC, other data will be made available once the classification is released by the ONS. The cluster outputs of the classification will be made available as a comma-separated values (CSV) file. It is a non-proprietary data format that can be read by multiple software packages, such as Microsoft Excel, SPSS and OpenOffice. This CSV file will contain all of the OAs and SAs for the UK and which Supergroup, Group and Subgroup they have been assigned to. With this CSV file a user will also be able to input the file into a geographic information system, such as ArcGIS or the open source alternative, Quantum GIS, and explore the classification for themselves. In these circumstances they would need to link CSV file to an ESRI shapefile, a file type used to represent geospatial vector data such as digital boundaries, which are made freely available by the respective statistical bodies for England and Wales, Scotland and Northern Ireland.

Perhaps the most important additional output from the 2011 OAC will be the R code that was developed. Once the classification has been released by the ONS, all of the code that was used in the creation of the 2011 OAC will be published to www.github.com/geogale/2011OAC. This will comprise of four main components:

1. The code used to convert, transform, standardise and cluster the dataset.
2. The code used to produce the numerous outputs of the clustering.
3. The code used to perform the WCSS analysis by clustering a dataset multiple times and the removal of a different variable each time
4. The code used to try and ascertain the number of clusters present within a dataset.

This should provide the basic framework for anyone wishing to either recreate the 2011 OAC or create a bespoke geodemographic classification of their own. The release of this code will be a key distinguishing factor between the 2001 OAC and the 2011 OAC. It makes the processes that were used more transparent and provides an opportunity for critical evaluation of one of the core components of the project.

7.5. Conclusion

The chapter has explained how the methodology outlined in Chapter 6 was applied in the creation of the 2011 OAC. The process of creating the 2011 OAC followed a number

of steps to guarantee an optimum result. Firstly, a number of initial variables to be considered for use in the classification were selected. This initial selection was subsequently significantly reduced to identify key variables to best describe the population characteristics of the UK, either by the removal or merging of variables. These decisions were made based on a combination of quantitative analysis and logical decision making processes to guarantee that the final variable selection was optimal for creating a good general purpose geodemographic classification.

The creation of a dataset comprising of the finalised set of variables led to the identification of the optimum methods of converting, transforming and standardising the dataset. A total of 27 different permutations were created from the three rate calculation techniques, three transformation techniques and three standardisation techniques used. Of these 27 datasets, 23 were discarded leaving four to undergo testing to identify the optimum method. Again, through the use of quantitative methods and logical deduction, a final dataset was identified as optimum. This led to the creation of a hierarchical classification structure, with 8 Supergroups, 26 Groups and 76 Subgroups being considered the ideal number of clusters to form the 2011 OAC. To complete the process of creating this new geodemographic classification, a number of outputs were produced. The descriptive and visual outputs provide the public interface to the 2011 OAC, distinguishing the 2011 OAC from being solely an output of cluster analysis, to being a fully functional geodemographic classification. An additional output was the proposed release of the R code that was used in the construction of the classification, contributing to the fully open nature of the geodemographic classification.

The creation of the 2011 OAC can be considered to be a success from a technical point of view. The methodology outlined in Chapter 6 functioned in both producing expected outputs which correlated with expectations, and in testing a more comprehensive set of factors in comparison to the methodology of the 2001 OAC. This does not however guarantee that the 2011 OAC will perform as expected. The overall success of the 2011 OAC can only really be judged by whether the users find it useful, and for this to happen it needs to meet their expectations. Expectations will vary between users, but a core component will be if the 2011 OAC names and pen portraits match a user's knowledge of an area. This basic premise of geodemographics would dictate that this will not happen for every OA and SA across the UK, but it should be expected that the 2011 OAC would match the reality of an area more often than not. As such, the evaluation of the 2011 OAC forms a core component of its construction. Assessment of different aspects

of the classification are discussed in Chapter 8; only after completion of this evaluation phase can the construction of the 2011 OAC be completed.

Chapter 8

Validation of the 2011 Area Classification for Output Areas

8.1. Introduction

The process for validating the 2011 Area Classification for Output Areas (2011 OAC) can be divided into categories: variable specification; cluster assignment certainty; homogeneity; changes between 2001 and 2011 and ground-truthing. Section 8.2 details how the final 60 variables used to construct the 2011 OAC performed when clustered. The use of within-cluster sum of squares (WCSS) analysis is used to determine the sensitivity of the 2011 OAC to individual variables. Section 8.3 explores the certainty of the cluster assignments to each Output Area (OA) and Small Area (SA) in the UK with the use of squared Euclidean distance (SED) values. This also allowed the propensity for each OA and SA to belong to the other Supergroups of the 2011 OAC to be assessed, highlighting the classification's fuzzy characteristics. Section 8.4 explores the further use of SED values to assess the level of homogeneity present within the 2011 OAC. Analysis was performed between the different hierarchical levels of the 2011 OAC, the different geographic areas and between individual Supergroups. This allowed for the identification of both areas in the UK and specific clusters, which contained more divergent population characteristics.

Section 8.5 details how changes that have occurred in the UK since 2001 have been incorporated into the 2011 OAC. A comparison of how change in the built environment along with social change over the past decade is performed by contrasting the 2011 OAC with the 2001 OAC in specific regions of the UK. Section 8.6 provides details of a ground-truthing exercise where participants visited certain areas in London to assess whether the assigned 2011 OAC Supergroup offered the best representation. Finally, Section 8.7 concludes by discussing the results of the validation exercises and implications for the overall robustness of the 2011 OAC and the consequences for users of the classification.

8.2. Variable specification

Lorenz curves and Gini Coefficients can be used to assess how well each variable performs in terms of categorising the UK's population. A Lorenz curve (Lorenz, 1905) is a commonly used graphical method of summarising the inequalities in the distribution of income and wealth (Arnold, 2008; Kakwani, 2010); and the Gini Coefficient (Gini, 1912) is a measure of statistical dispersion.

Lorenz curves and Gini Coefficients were repurposed to analyse the disparity between the assignment of a variable to each of the Output Areas (OAs) and Small Areas (SAs) across the UK. Figure 8.1 shows the Lorenz curves for each of the 60 variables that were used to construct the 2011 OAC. The diagonal black line represents an equal distribution of variables across all of the OAs and SAs in the UK. The values on the X-axis represent the percentage of OAs and SAs that have been assigned to each variable, and the values in the Y-axis are the total percentage of OAs and SAs in the UK. The closer the curve is to this black line, the greater equality that variable demonstrates in terms of representation across the whole of the UK.

Figure 8.1 provides a clear visual indication of the level of disparity that exists between the 2011 OAC variables. The variables 'Persons who are white' and 'Employed persons aged between 16 and 74 who work full-time' are the two most evenly distributed variables across the UK. It is indicated that nearly 100% of the UK's OAs and SAs contain persons who correspond to these variables. Conversely, 'Persons living in a communal establishment' is the most unevenly distributed variable in the 2011 OAC, with 100% of the variable's occurrence being found in less than 10% of the UK's OAs and SAs. Other notable examples of variables that have unequal distributions are 'Persons who are Asian/Asian British: Bangladeshi' along with the Pakistani equivalent, 'Households with full-time students' and 'Employed persons aged between 16 and 74 who work in the agriculture, forestry or fishing industries'.

Gini coefficients provide a summary statistic of the equality of distribution for each variable. A lower Gini Coefficient represents a more equal distribution of the variable; a value of 0 would indicate complete equality in the distribution of a variable across the UK, with values closer to 1 indicating an increasing level of inequality in their distribution. Figure 8.2 plots the Gini Coefficients for each of the 2011 OAC's 60 variables.

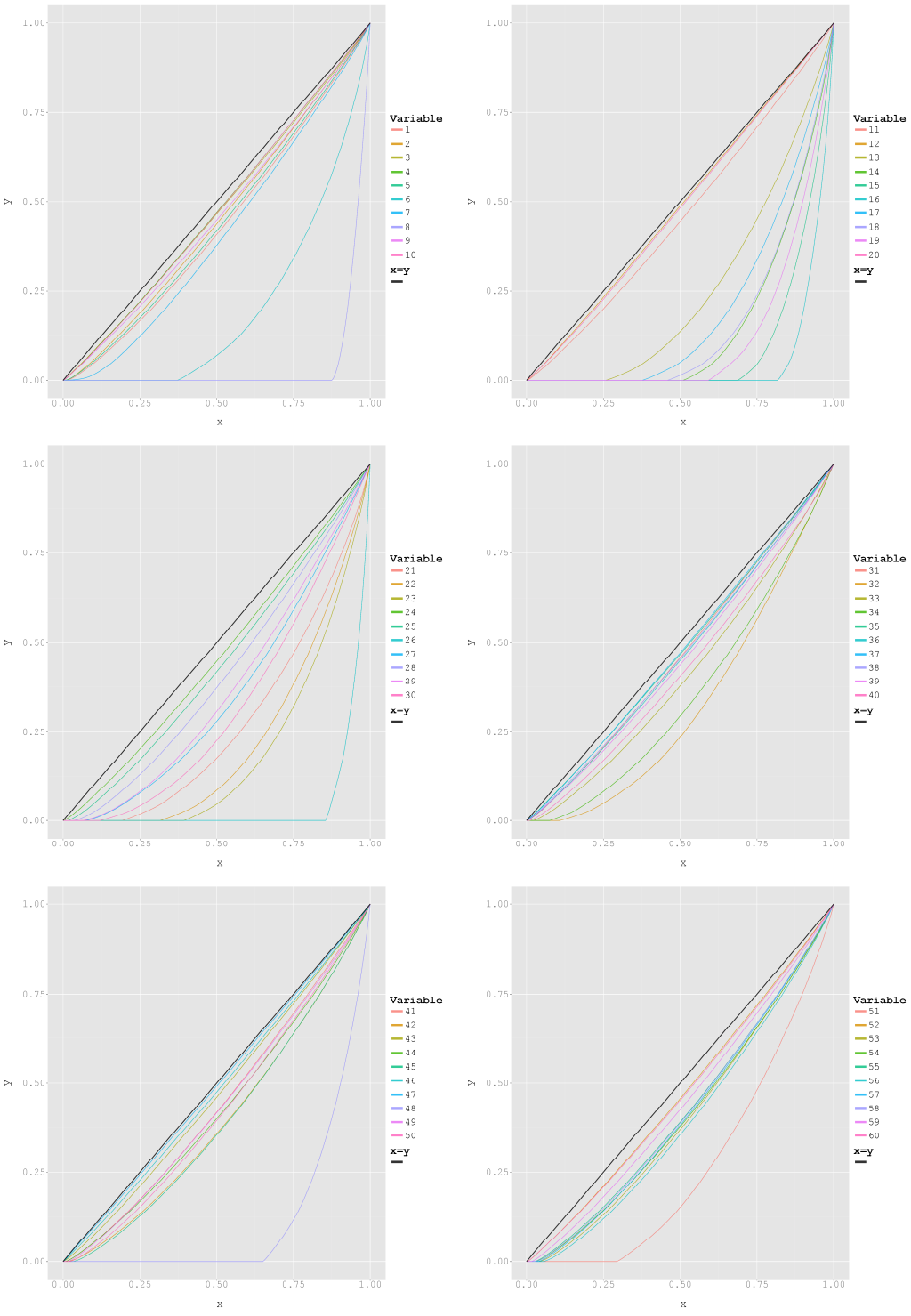


Figure 8.1: Lorenz curves for the 2011 OAC variables

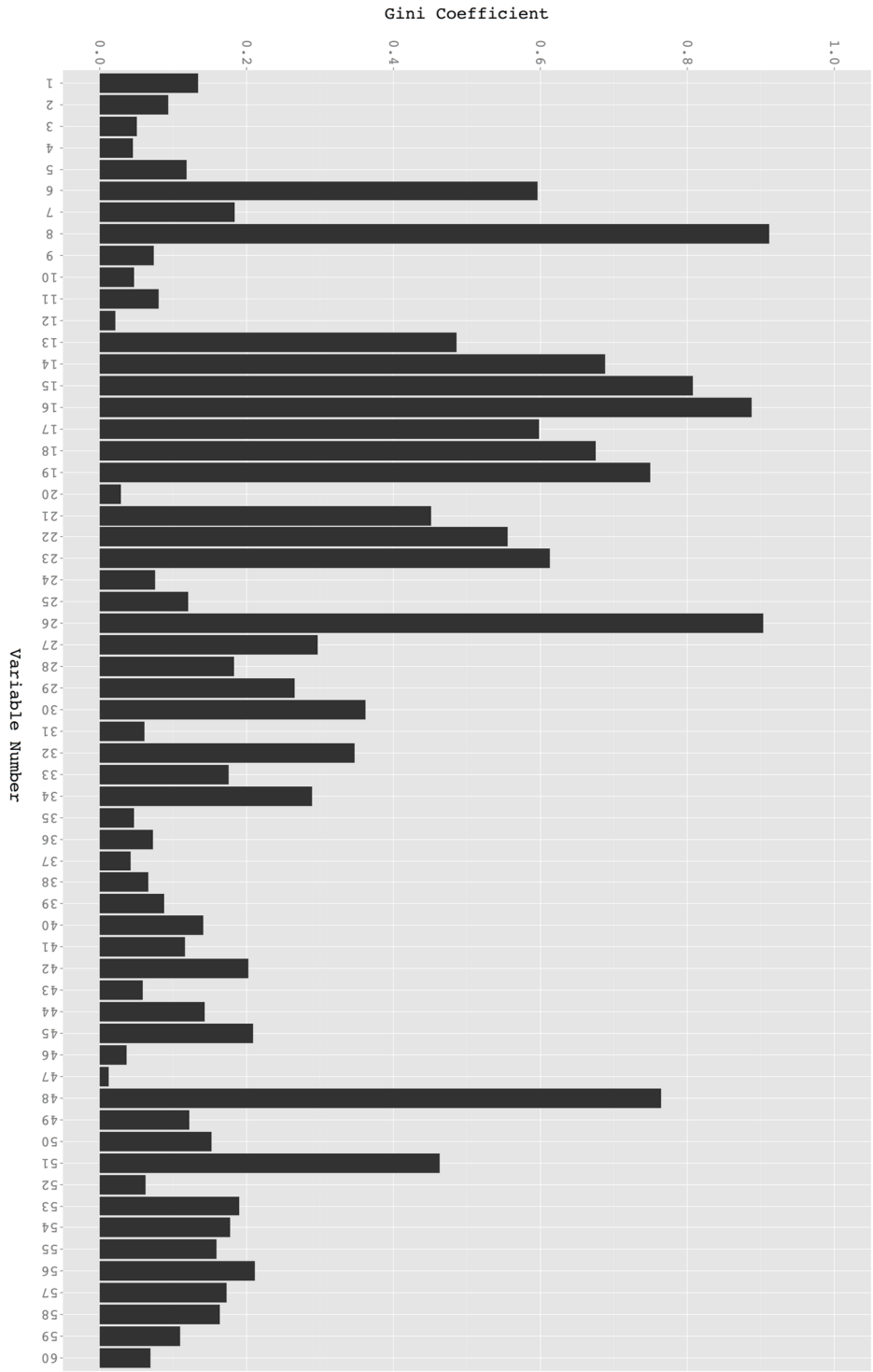


Figure 8.2: Gini Coefficients for the 2011 OAC variables

Gini coefficients offer a quantitative method for comparison of the distribution of variables that is not achievable through the visual representation provided by the Lorenz curves. The mean Gini Coefficient for the 2011 OAC variables is 0.27. Out of the total 60 variables, 41 have a below average level of inequality in their distributions. The results for each variable are shown in Table D.1 in Appendix D.

The distribution of variables across the UK is of significance for the interpretation of the 2011 OAC clusters. All variables have complete coverage of the UK, albeit in some instances with zero values. As such, not every OA or SA in the UK is guaranteed to contain a value above zero for each of the variables used to construct the 2011 OAC. The mean Gini Coefficient of 0.27 indicates that, on average, the total occurrences of a variable in the UK are found in 73% of the OAs and SAs. This can cause certain variables to be misleading in their comparison with other variables in a cluster. For example, variables with high Gini Coefficients, such as 'Persons living in a communal establishment', with a value of 0.91, will only represent 9% of OAs and SAs in the UK. In comparison, the 'Employed persons aged between 16 and 74 who work full-time' variable has a Gini Coefficient of 0.01, meaning it is represented in 99.9% of the UK's OAs and SAs.

The existence of zero values for the majority of variables results in an inconsistency in the 'national mean' for each variable, which is particularly prominent at the Supergroup level. The distribution of a variable such as, 'Households with full-time students', for example, is more likely to be spatially clustered around universities in the UK. As such, the mean of this variable does not reflect the geographical unevenness of its distribution across the UK, but rather the limited geographic locations where this variable is distributed. Therefore, comparison of this 'national mean' with that of a variable with more even geographic coverage such as 'Persons who are white', does not reflect the contrasting coverage and the subsequent differences in the UK's population represented by each.

There is no way to distinguish whether a cluster contains a below average or a zero value of a variable. The zero values on radial plots for example are relative to a mean value, and not necessarily an indication that clusters do not contain a particular variable. The Lorenz curves and Gini Coefficients of the 2011 OAC variables allow for the identification of variables that are likely to exhibit a highly uneven distribution. However, the visual and statistical overview provided by Lorenz curves and Gini Coefficients are insufficient to identify the geographic location and concentration of variables. For interpretative

purposes it is important to determine whether variables are unevenly spatially clustered or are distributed more equally across the UK. It is therefore necessary to utilise a tool such as the '2011 Census Open Atlas' created by Dr Alex Singleton (and available from www.alex-singleton.com/Open-Atlas/), which provides a convenient method of mapping the distributions of the variables. In conjunction with the Lorenz curves and Gini Coefficients, a detailed logical analysis of the specification of the 2011 OAC variables can be completed.

Although the Lorenz curves and Gini Coefficients do not incorporate a spatial component into their outputs, they allow differences in the 2011 OAC variable distributions to be identified. This allows each variable to be individually assessed on the impact it has on the creation of the 2011 OAC. These techniques provide additional insight into the formation of the classification, and how differences in variable distribution impact interpretation of the final clusters.

In addition to the Lorenz curves and Gini Coefficients, the within-cluster sum of squares (WCSS) analysis technique (as described in Section 6.6.3), can be used to assess the performance of the 2011 OAC variables when clustered in the final solution. Figure 8.3 shows the result of the WCSS analysis on the 2011 OAC variables. It indicates that the classification is particularly sensitive to variables in the 'Housing' domain, and five in particular:

- 'Households who live in a flat'
- 'Households who live in a terrace or end-terrace house'
- 'Households who are social renting'
- 'Households who live in a detached house or bungalow'
- 'Households who live in a semi-detached house or bungalow'

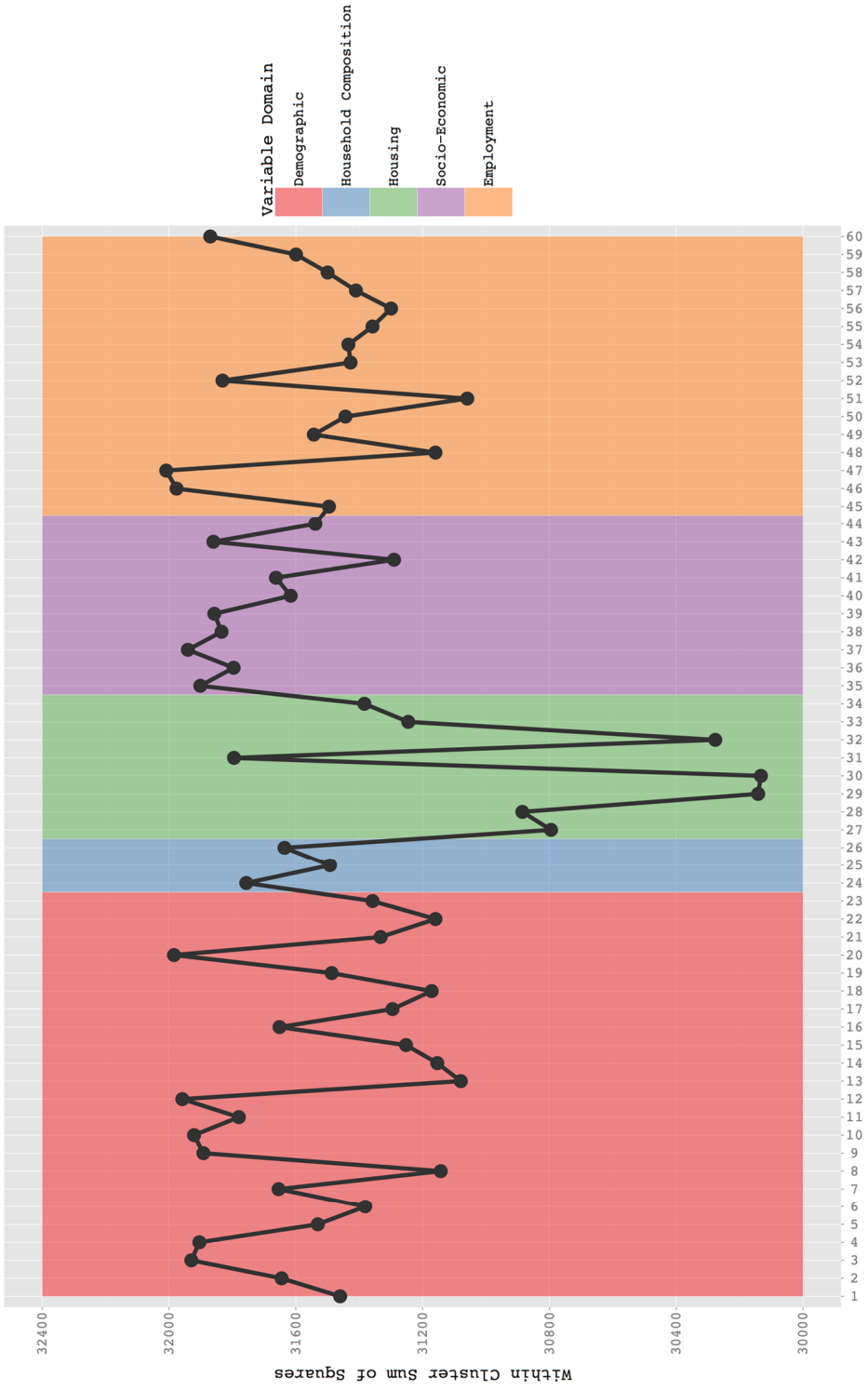


Figure 8.3: WCSS analysis on the 2011 OAC Variables
(See Table 7.3 for variable names)

Similar to reasons discussed in Section 7.2.2.2, housing variables appear to be more susceptible to adversely impacting cluster homogeneity when clustered as part of a large multivariate dataset. The 2011 OAC is likely to be sensitive to the five variables identified because of the way in which they interact with other variables. The simplistic nature of categorising types of housing into a small number of groups has an impact on their interaction with other characteristics of the population. The characteristics of different sections of the population in the UK may differ greatly, but they can become linked when they are prevalent within the same type of housing. As a result, the final clusters of the 2011 OAC are less homogenous than if these five variables had been excluded. Additionally, the widely divergent characteristics of a housing type make it difficult for a clustering algorithm to identify commonality with other variables.

Aside from those in the 'Housing' domain, the variation in WCSS value demonstrated for the remainder of the 2011 OAC variables is less severe. This suggests that whilst each variable does have an impact on the overall homogeneity of the final clusters, their presence is less significant on the 2011 OAC. It can therefore be concluded that from a clustering perspective, the variables have performed well.

Evidence from the Lorenz curves, Gini Coefficients and WCSS analysis suggests that whilst the distributions of the 60 selected variables may have differed and although some of the clusters were not as homogenous as they could have been, the result of their interactions across the UK led to the distinct nature of the 8 Supergroups, 26 Groups and 76 Subgroups discussed in Section 7.4. The clustered variables have additionally allowed for unique names and pen portraits to be formed for each cluster. This suggests that they performed well in helping to create a unique and dynamic clustering solution. The overall performance of the variables is therefore linked to the overall performance of the 2011 OAC, which cannot be assessed until it has been released and actively used.

8.3. Cluster assignment certainty

The assignment of any OA or SA in the UK to one of the 8 Supergroups, 26 Groups or 76 Subgroups is based on the squared Euclidean distance (SED) value calculated as part of the k-means clustering algorithm. As discussed in Section 6.2, the SED is a dissimilarity measure; the larger the SED value for each OA or SA, the more dissimilar it is to the cluster centroid (the average characteristics of that cluster's population). Therefore, the Supergroup assigned to each OA or SA is determined by the smallest SED value.

The extent to which the SED values for the Supergroups assignments differ provide a proxy measure of uncertainty. The reasoning behind the assignment of a Supergroup to an OA or SA can be split into two categories. Firstly, the assignment offered the best representation. For example, in a rural location the resident population is so distinct from residents in urban areas, that the 'Rural Residents' Supergroup is likely to be the only assignment with a low SED value. As a consequence there is greater certainty in the cluster assignment.

Secondly, a Supergroup will be assigned to an OA or SA because it offers greater benefits than those of other assignment options. This rationale is more likely to apply to urban areas, where within a single OA or SA a great deal of variation in the population characteristics can be found. This variation causes difficulties in clustering algorithms as an area does not obviously belong to any one pre-existing cluster. Instead, the area is assigned to the least bad option. The resulting SED value will therefore be higher, indicating that the characteristics of the area conform less to the cluster average. This leads to an increased possibility that characteristics of the area share commonalties with other Supergroups, which causes distinctions between them to be less significant.

Figure 8.4 is an equalising density cartogram map of the SED values for the OAs and SAs assigned to each 2011 OAC Supergroup in the UK, with each OA and SA reflecting their total population in 2011. The darker the colour the higher the level of uncertainty represented. A distinct spatial pattern is revealed; notably it appears that the Supergroup assignment for Scotland is a great deal more uncertain in comparison to the rest of the UK. On average, the mean SED value for the OAs assigned to Supergroups in Scotland is 1.13, compared to 0.86 for the rest of the UK. There are pockets of greater uncertainty across other parts of the UK, notably in urban areas, but these are on a smaller scale compared to Scotland. This disparity in uncertainty found in Scotland is due to the comparatively large size of England's population in comparison to the rest of the UK. England has 84% of the UK's population, which results in the characteristics of the 2011 OAC clusters being predominantly based on the characteristics of the English population. Scotland, Wales and Northern Ireland are therefore classified on the basis of how similar they are to the average English characteristics. The greater certainty in the cluster assignments for Wales and Northern Ireland implies that they share more similar population dynamics with England than Scotland does.

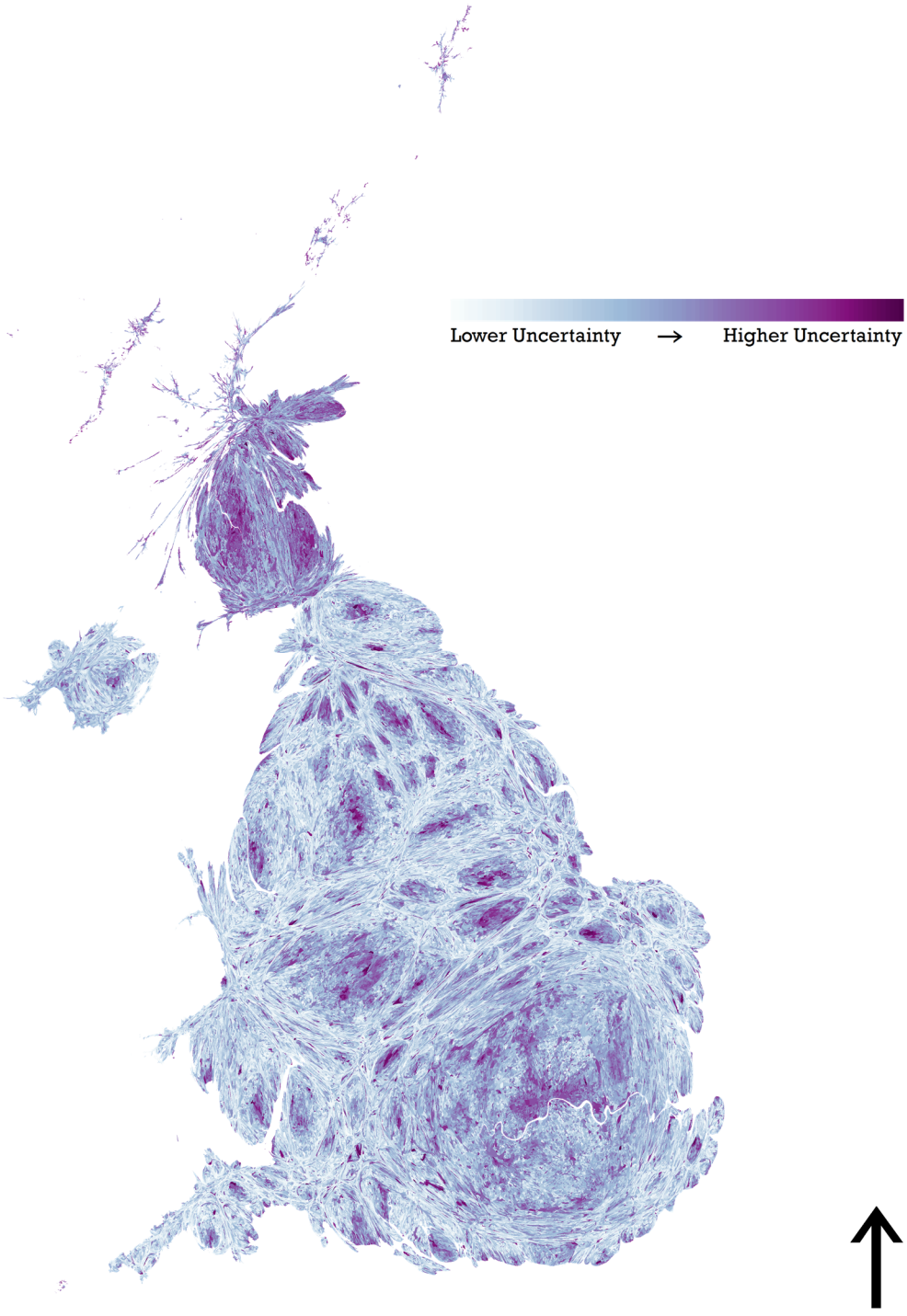


Figure 8.4: Certainty of the 2011 OAC Supergroup assignment across the UK

This greater uncertainty in Scotland is nationwide, but it is notable that the rural areas of the country are distinctly different to equivalent areas in the rest of the UK. This conclusion can be reached by looking at the fuzzy characteristics of the classification. As discussed in Section 6.8.1.5, the concept of fuzzy classification is based on an area having proportional membership to all of the final clusters, rather than belonging to only one. Although the 2011 OAC is not explicitly a fuzzy classification, utilisation of the k-means clustering processes permits the exploration of the 2011 OAC's fuzzy characteristics.

In addition to assessment of the extent to which an OA or SA contains the characteristics of its assigned Supergroup, the SED value also assesses alternative Supergroups. Figures D.1 to D.8 in Appendix D visualise the propensity for each OA and SA in the UK to conform to each of the eight 2011 OAC Supergroups. The darker shading suggests that those areas conform more closely to the pen portraits of each cluster (see Appendix C). Figure D.1 clearly shows that the areas in Scotland conform less to the average characteristics of the 'Rural Residents' Supergroup. However, the rural areas in Scotland have a greater similarity to this Supergroup than any of the alternatives, hence why they are assigned to it. The extent to which Scotland's rural areas actually differ to the remainder of the UK, is a result of the differences in the creation of the OA geography (as discussed in Section 8.4). Assignment to OAs in Scotland is therefore more uncertain.

Whilst visually prevalent in Figure 8.4, the level of uncertainty seen in Scotland for the assignment of Supergroups, represents a relatively small proportion of the UK's population. The smaller population size of Scotland (just over 8% of the UK's total), is not large enough to fundamentally change the average characteristics of the Supergroup for the whole of the UK. If the 'Rural Residents' Supergroup were to be made more representative for Scotland, it would increase levels of uncertainty for the remainder of the UK.

The larger the geographic extent of a geodemographic classification, the greater the number of unique areas covered as part of the process. This leads to a decrease in the certainty of the cluster assignment as it becomes increasingly difficult to find commonalities between the different locations. Utilisation of the SED values allow these differences and inherent uncertainties to be identified, and the extent to which areas that do not share the same cluster assignment are similar can be recognised. This increased knowledge of an area or region can provide a greater context into how the dynamics at the small area level influence the whole classification.

The research undertaken by Slingsby et al. (2011) analysing SED values from the 2001 OAC, concluded that certain areas in the UK shared similarities with multiple Supergroups, whilst others only had the dominate characteristics of one Supergroup. This pattern can also be identified within the 2011 OAC. The 'Constrained City Dwellers' Supergroup in London is only assigned to 1.12% of the capital's OAs. Yet based on Figure D.7, it would appear that a larger number of OAs actually share similar characteristics, especially towards the east of the city. A further example can be demonstrated by the 'Ethnicity Central' Supergroup in Glasgow. It is only assigned to 7.27% of the city, yet when viewing Figure D.3, it becomes clear that a large number of OAs in the Glasgow region share many characteristics with the Supergroup.

The examples from London and Glasgow can be seen visually at a national level. Similar patterns are also likely to be prevalent at the neighbourhood level, providing a greater understanding to the dynamics of the 2011 OAC at this scale. Without exploring the fuzzy characteristics of the classification, such patterns would go unnoticed. The SED values allow areas of greater uncertainty in a cluster assignment to be identified. The values can also demonstrate that the complexity of a geodemographic classification can be lost when each area is only assigned a single simplified cluster. SED values additionally allow for the creation of a de facto fuzzy classification. This ultimately provides a better knowledge of the population dynamics of the UK.

8.4. Homogeneity of the 2011 OAC

Section 8.3 detailed how SED values calculated for individual OAs or SAs can be utilised for analysis of the relative centrality of their cluster assignments. However, instead of detailing the certainty of the 2011 OAC as a whole, the classification can be analysed from a number of different perspectives to examine the variation that exists in this certainty. These include exploration of the variations between: the three different hierarchical levels of the classification; individual clusters; and geographic areas.

8.4.1. Homogeneity between hierarchical levels

An aim of the 2011 OAC was to have as homogenous clusters as possible whilst providing the best geographical representation across the UK. In total, each OA and SA in the UK was assigned three SED values, one for each level of the 2011 OAC hierarchy. Table 8.1

details the statistical overview of the variation in SED values across each of the three hierarchical levels. The maximum and minimum SED values recorded remain consistent across the different hierarchical levels. There is also consistency in the number of OAs and SAs with values above and below the respective mean SED values. The frequency of the SED values calculated for each OA and SA is shown in Figure 8.5. The majority of SED values for all three levels of the hierarchy range from 0.4 to 1.6. A long tail can however be identified for each of the hierarchies, with each distribution being positively skewed, indicating that there are areas with comparatively high SED values.

The average cluster homogeneity does however increase as you move down the 2011 OAC hierarchy. While this is an expected result, as discussed in Section 7.3.1, it does not signify that the level of homogeneity between individual clusters remains consistent.

Table 8.1: 2011 OAC hierarchy SED values overview

	Supergroups	**Groups**	**Subgroups**
Maximum SED value	3.16	3.06	2.98
Minimum SED value	0.37	0.35	0.30
Mean SED	0.91	0.85	0.80
Number of OAs and SAs below mean	57.3% (133,054)	57.5% (133,448)	58.2% (135,156)
Number of OAs and SAs above mean	42.7% (99,242)	42.5% (98,848)	41.8% (97,140)

(Counts are in brackets)

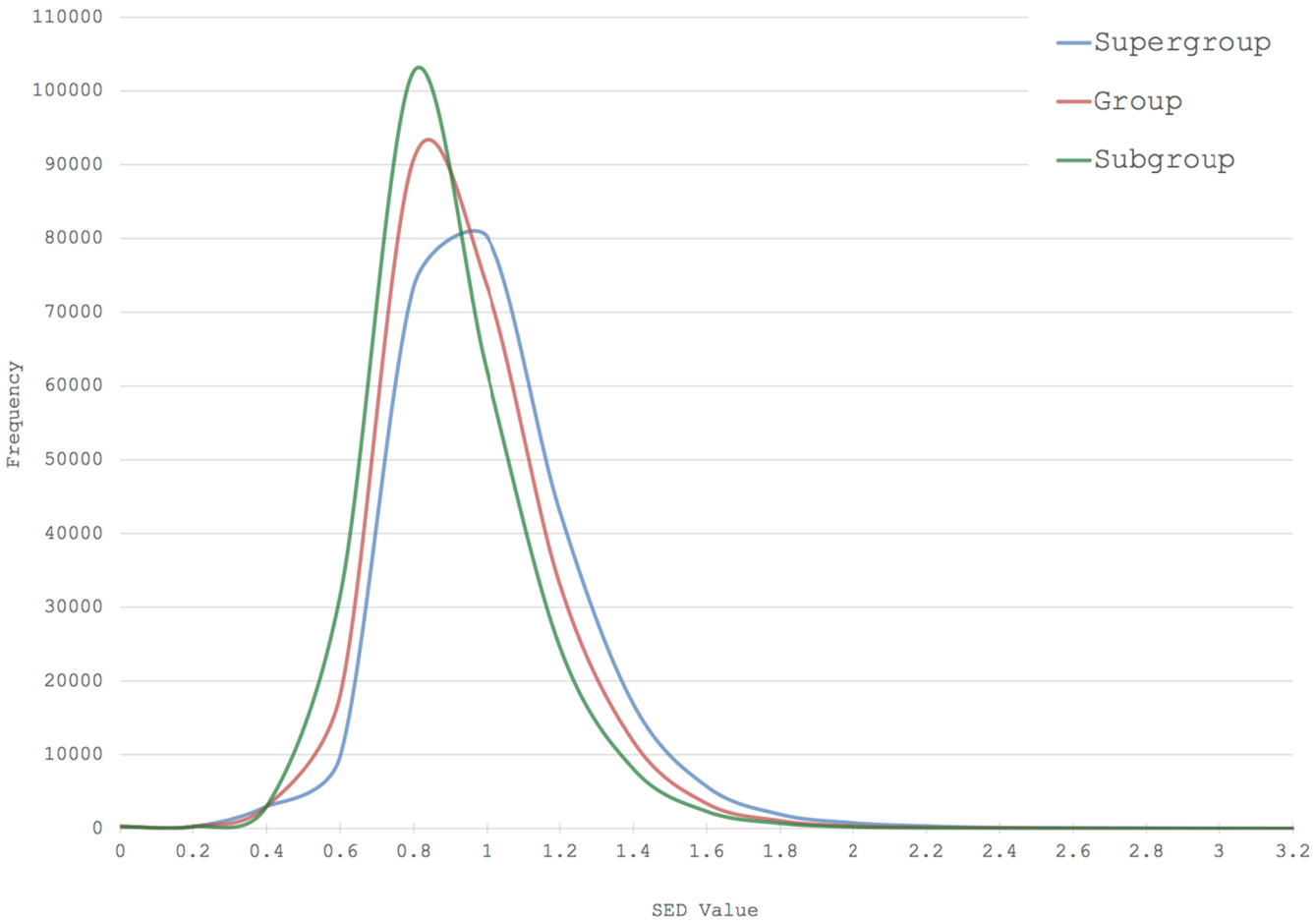


Figure 8.5: Frequency of 2011 OAC hierarchy SED values

Table 8.2 details the SED values for each the 2011 OAC clusters. The values indicate that the 'Rural Residents' Supergroup is the most homogenous, despite the disparity with areas in Scotland, as discussed in Section 8.3. The 'Cosmopolitans' Supergroup has the lowest level of homogeneity, an indication that there is greater variance in the population characteristics in the areas classified as this cluster. As discussed in Section 7.3.2, although the values for each Supergroup can be directly compared, the values of the Groups and Subgroups cannot be. Groups and Subgroups can only be compared to each other when they are formed from the same parent Supergroup or Group. Although it is not possible to compare all 26 Groups or 76 Subgroups together, the homogeneity of clusters at these levels derived from their parent Supergroup can be analysed. The SED values should decrease as you move down the hierarchy, although there are some exceptions to this. In total 6 Groups and 16 Subgroups have higher SED values than their parent Supergroup or Group, suggesting a small proportion of clusters are less homogenous in the middle and lower tiers of the 2011 OAC.

Table 8.2: 2011 OAC SED values per cluster

Supergroup	SED	Group	SED	Subgroup	SED
1 - Rural Residents	0.82	1a - Farming Communities	0.79	1a1 - Rural Workers and Families	0.80
				1a2 - Established Farming Communities	0.66
				1a3 - Agricultural Communities	0.74
				1a4 - Older Farming Communities	0.84
		1b - Rural Tenants	0.74	1b1 - Rural Life	0.69
				1b2 - Rural White-Collar Workers	0.66
				1b3 - Ageing Rural Flat Tenants	0.78
		1c - Ageing Rural Dwellers	0.85	1c1 - Rural Employment and Retirees	0.88
				1c2 - Renting Rural Retirement	0.75
				1c3 - Detached Rural Retirement	0.78
2 - Cosmopolitans	1.19	2a - Students Around Campus	1.07	2a1 - Student Communal Living	1.12
				2a2 - Student Digs	0.94
				2a3 - Students and Professionals	0.91
		2b - Inner-City Students	1.15	2b1 - Students and Commuters	1.14
				2b2 - Multicultural Student Neighbourhoods	1.05
		2c - Comfortable Cosmopolitans	1.12	2c1 - Migrant Families	0.97
				2c2 - Migrant Commuters	1.12
				2c3 - Professional Service Cosmopolitans	1.09
		2d - Aspiring and Affluent	0.91	2d1 - Urban Cultural Mix	0.88
				2d2 - EU White-Collar Workers	0.78
				2d3 - Highly-Qualified Quaternary Workers	0.86
3 - Ethnicity Central	0.98	3a - Ethnic Family Life	0.86	3a1 - Established Renting Families	0.81
				3a2 - Young Families and Students	0.83
		3b - Endeavouring Ethnic Mix	0.85	3b1 - Striving Service Workers	0.73
				3b2 - Bangladeshi Mixed Employment	0.77
				3b3 - Multi-Ethnic Professional Service Workers	0.84

Supergroup	SED	Group	SED	Subgroup	SED
3 - Ethnicity Central	0.98	3c - Ethnic Dynamics	1.12	3c1 - Constrained Neighbourhoods	1.00
				3c2 - Constrained Commuters	1.23
		3d - Aspirational Techies	0.80	3d1 - Established Tech Workers	0.80
				3d2 - Old EU Tech Workers	0.72
				3d3 - New EU Tech Workers	0.72
4 - Multicultural Metropolitans	0.97	4a - Rented Family Living	0.87	4a1 - Social Renting Young Families	0.80
				4a2 - Private Renting New Arrivals	0.87
				4a3 - Commuters with Young Families	0.77
		4b - Challenged Asian Terraces	0.92	4b1 - Asian Terraces and Flats	0.83
				4b2 - Pakistani Communities	0.91
		4c - Asian Traits	0.90	4c1 - Achieving Minorities	0.87
				4c2 - Multicultural New Arrivals	0.75
				4c3 - Inner City Ethnic Mix	0.84
5 - Urbanites	0.91	5a - Urban Professionals and Families	0.85	5a1 - White Professionals	0.74
				5a2 - Multi-Ethnic Professionals with Families	0.81
				5a3 - Families in Terraces and Flats	0.84
		5b - Ageing Urban Living	0.91	5b1 - Delayed Retirement	0.92
				5b2 - Communal Retirement	0.89
				5b3 - Self-Sufficient Retirement	0.78
6 - Suburbanites	0.84	6a - Suburban Achievers	0.86	6a1 - Indian Tech Achievers	0.82
				6a2 - Comfortable Suburbia	0.79
				6a3 - Detached Retirement Living	0.76
				6a4 - Ageing in Suburbia	0.83
		6b - Semi-Detached Suburbia	0.77	6b1 - Multi-Ethnic Suburbia	0.78
				6b2 - White Suburban Communities	0.69
				6b3 - Semi-Detached Ageing	0.70
				6b4 - Older Workers and Retirement	0.73
7 - Constrained City Dwellers	1.07	7a - Challenged Diversity	0.88	7a1 - Transitional Eastern European Neighbourhoods	0.90
				7a2 - Hampered Aspiration	0.83
				7a3 - Multi-Ethnic Hardship	0.78
		7b - Constrained Flat Dwellers	1.17	7b1 - Eastern European Communities	1.10
				7b2 - Deprived Neighbourhoods	1.22
				7b3 - Endeavouring Flat Dwellers	1.06
		7c - White Communities	0.93	7c1 - Challenged Transitionaries	0.88
				7c2 - Constrained Young Families	0.91
				7c3 - Outer City Hardship	0.87
		7d - Ageing City Dwellers	1.21	7d1 - Ageing Communities and Families	1.02
				7d2 - Retired Independent City Dwellers	1.17
				7d3 - Retired Communal City Dwellers	0.98
				7d4 - Retired City Hardship	1.64
8 - Hard-Pressed Living	0.83	8a - Industrious Communities	0.75	8a1 - Industrious Transitions	0.69
				8a2 - Industrious Hardship	0.76
		8b - Challenged Terraced Workers	0.82	8b1 - Deprived Blue-Collar Terraces	0.73
				8b2 - Hard-Pressed Rented Terraces	0.83
		8c - Hard-Pressed Ageing Workers	0.76	8c1 - Ageing Industrious Workers	0.69
				8c2 - Ageing Rural Industry Workers	0.79
				8c3 - Renting Hard-Pressed Workers	0.73
		8d - Migration and Churn	0.77	8d1 - Young Hard-Pressed Families	0.72
				8d2 - Hard-Pressed Ethnic Mix	0.71
				8d3 - Hard-Pressed European Settlers	0.74

Decreasing homogeneity as you move down the hierarchies of the 2011 OAC is undesirable. An example shown in Table 8.2 from the 'Constrained City Living' Supergroup indicates that the 'Ageing City Dwellers' and 'Flat Dwellers' Groups are less homogenous than their parent Supergroup. In addition, the 'Retired City Hardship' and 'Deprived Neighbourhoods' Subgroups are even less homogenous than their respective parent Group. This has occurred in these clusters due to the large variation in the characteristics of the resident population. Detection of any consensus within these clusters is difficult, and identification of these specific characteristics would involve a more in-depth knowledge of an area which statistics alone cannot provide.

In other instances, a decrease in the SED values as you move down the hierarchy is an indication of the clusters becoming increasingly distinctive as fewer OAs and SAs get assigned to them. However, clusters with higher SED values than their parent Supergroup or Group are not necessarily less representative. Although there will be greater variation within the population, the names and pen portraits attached to these clusters will still provide a better representation of each area when compared to their parent Supergroup or Group.

Analysis of the SED values as you move down the classification hierarchy provides evidence to identify the extent to which residents classified into any particular cluster will vary from the 'average' characteristics of the cluster. This can be further developed to look at the geographical distribution of homogeneity within clusters and identification of any regions or countries in the UK that are more likely to have OAs or SAs assigned to atypical 2011 OAC Supergroups.

8.4.2. Homogeneity between clusters and geographical areas

For the purpose of analysing cluster homogeneity, any OA or SA with a SED value of over 1.5 for their Supergroup assignment is considered to be an outlier. This threshold value was also used for a similar study with the 2001 OAC (Vickers, 2006). Table 8.3 details the number of OAs and SAs that can be considered as outliers for the different regions in England and Wales, Scotland and Northern Ireland. On average, 1 in every 44 of the UK's OAs and SAs can be considered to be an outlier. This concentration does vary between locations, from 1 in every 13 in Scotland, to 1 in every 349 in Northern Ireland. Across the UK, 66.3% of the OAs and SAs classified as outliers are found in Scotland, compared to London which only accounts for 3.2%.

Table 8.3: 2011 OAC outliers by English Regions, Wales, Scotland and Northern Ireland

English Regions, Wales, Scotland and Northern Ireland	Count of Outliers	Percentage of Outliers	Supergroup Total	Percentage of Supergroups are Outliers	Mean SED per OA or SA	Mean Population per OA or SA
East Midlands	208	3.94%	14,706	1.41	308	0.86
East of England	152	2.88%	18,995	0.80	308	0.84
London	168	3.18%	25,053	0.67	326	0.96
North East	73	1.38%	8,802	0.83	295	0.82
North West	251	4.75%	23,343	1.08	302	0.85
Northern Ireland	13	0.25%	4,537	0.29	399	0.83
Scotland	3,500	66.28%	46,351	7.55	114	1.13
South East	275	5.21%	27,638	1.00	312	0.85
South West	160	3.03%	17,644	0.91	300	0.83
Wales	68	1.29%	10,036	0.68	305	0.82
West Midlands	174	3.29%	17,916	0.97	313	0.87
Yorkshire and the Humber	239	4.53%	17,275	1.38	306	0.86
UK	5,281	100%	232,296	2.27	272	0.91

The disparity between Scotland and London is of particular interest due to the unique status of London within the UK (Sassen, 2001). This would suggest that the resident population of London would be equally unique and therefore less likely to conform to the average 2011 OAC Supergroup characteristics. The reasoning behind this disparity is a result of two factors. Firstly, a different methodology was utilised in the creation of the Census geography in Scotland in comparison to the remainder of the UK. As discussed in Section 3.3.1, the design of OAs in Scotland was focussed on ensuring that urban and rural locations were not mixed together, rather than ensuring social homogeneity, as was applied to other areas of the UK. Additionally, the application of different minimum thresholds for population and households generated more OAs in Scotland than if the methodology used for the remainder of the UK had been applied.

The difference in methodologies results in Scotland's 8.38% of the UK's population represented as 19.95% of the UK's OA and SAs. This design of areal units in Scotland therefore exacerbates the number of areas classified as outliers. Population characteristics confined to a single OA or SA elsewhere in the UK instead span multiple

areal units. As these are less socially homogenous in Scotland it therefore makes it more likely they will be classed as outliers.

Secondly, Scotland contains more outliers than London due to the composition and size of its population. The size of London's population compared to Scotland's (8.2 million in comparison to 5.3 million), impacts how the clustering algorithm partitions the data. Due to the greater population concentration in London areas containing characteristics, that would elsewhere be classified into pre-existing clusters as outliers, were instead clustered together to create London-focused Supergroups. For example, 78% of the OAs and SAs classified into the 'Ethnicity Central' Supergroup are located in London. As such, London contains such a variation in population characteristics, that the classification is forced to accommodate this through the creation of a London focussed Supergroup.

Identifying the areal units which contain outliers provides a useful perspective on the certainty of the 2011 OAC Supergroup assignment. It is however also useful to understand which Supergroups are more likely to contain outliers. Table 8.4 provides a breakdown of the 5,281 OAs and SAs classed as outliers by their Supergroup assignment. Two Supergroups, 'Constrained City Dwellers' and 'Cosmopolitans' contain 63.5% of all outliers in the UK. The SED values would suggest that these are the least homogenous Supergroups, so it can be expected that they contain the most outliers. Table 8.5 shows a cross-tabulation of the outlier results by region/country and Supergroup. It illustrates that 84.4% of OAs or SAs classified as 'Constrained City Dwellers' outliers are located in Scotland. As already discussed, there is a greater chance of an OA being classed as an outlier in Scotland due to the areal unit design. Indeed, with the exception of 'Multicultural Metropolitans', all other Supergroups have the highest count of outliers in Scotland. This Supergroup is the exception as 'Multicultural Metropolitans' only represents 0.38% of OAs in Scotland. However, only 50% and 51% of outliers in the 'Cosmopolitans' and 'Urbanites' Supergroups are found in Scotland. This would suggest that there are particular population characteristics identified by these clusters that make them less homogenous. However, as around half the remaining outliers are distributed across the rest of the UK it is harder to suggest with any certainty what these characteristics may be. Detailed analysis of these remaining areas is impractical due to the large number involved. Although the utilisation of SED values the analysis of individual Supergroup outliers and the geographic variations that exist within these are useful, these methods are unable to explain why such disparity exists.

Table 8.4: 2011 OAC outliers by Supergroup

Supergroup	Count of Outliers	Percentage of Outliers	Supergroup Total	Percentage of Supergroups are Outliers	Mean SED per OA or SA	Mean Population per OA or SA
Rural Residents	301	5.70	27,300	1.10	269	0.82
Cosmopolitans	1,379	26.11	13,125	10.51	241	1.19
Ethnicity Central	369	6.99	11,849	3.11	303	0.98
Multicultural Metropolitans	223	4.22	23,502	0.95	350	0.97
Urbanites	530	10.04	38,697	1.37	286	0.91
Suburbanites	322	6.10	46,850	0.69	280	0.84
Constrained City Dwellers	1,974	37.38	27,135	7.27	184	1.07
Hard-Pressed Living	183	3.47	43,838	0.42	266	0.83
UK	**5,281**	**100**	**232,296**	**2.27**	**272**	**0.91**

Table 8.5: 2011 OAC outliers cross-tabulated by English Regions, Wales, Scotland and Northern Ireland and Supergroup

English Regions, Wales, Scotland and Northern Ireland / Supergroups	Rural Residents	Cosmopolitans	Ethnicity Central	Multicultural Metropolitans	Urbanites	Suburbanites	Constrained City Dwellers	Hard-Pressed Living
East Midlands	15	84	12	29	16	14	30	8
East of England	11	30	6	16	34	25	26	4
London	0	100	48	10	5	2	3	0
North East	2	37	5	1	2	3	21	2
North West	2	97	12	50	13	13	60	4
Northern Ireland	0	3	0	0	6	2	2	0
Scotland	215	693	261	25	271	217	1,665	153
South East	19	83	4	21	80	15	52	1
South West	15	41	0	7	56	12	28	1
Wales	1	39	1	1	11	3	11	1
West Midlands	16	64	14	24	12	9	32	3
Yorkshire and the Humber	5	108	6	39	24	7	44	6
UK	301	1,379	369	223	530	322	1,974	183

There are 103 OAs and SAs considered extreme outliers (those with SED values greater than 2.5) at Supergroup level for the 2011 OAC. To determine why they exist, analysis of these areas was performed using a combination of satellite imagery and Google Street View. It was recognised that in a number of cases, the OA contained either a mixture of residential and industrial areas or residential and communal establishments. It was necessary for residential and industrial areas to be combined within an OA or SA to ensure the 2011 OAC gave continuous coverage across the UK. Consequently, locations which have no night-time population, such as industrial estates, are grouped in with residential areas. It is therefore likely that multiple residential areas are incorporated with different population characteristics. Similarly, when residential and communal establishments are combined, divergent population characteristics are grouped together within a single OA or SA. Figure 8.6 is an example of this, with a care home in Plymouth being in the same OA as residential properties. The lack of social homogeneity that exists in this OA and others like it led to these locations being classed as extreme outliers.

There are however significant limitations to the use of aerial imagery and still photographs to analyse an area. Primarily, the evaluation is based on the built environment of an area. This is illustrated in Figure 8.7, an example of an extreme outlier found in an OA in Wokingham, a town in the English county of Berkshire. Based on the satellite imagery the OA appears to consist of a housing estate which would not look out of place in other towns and cities elsewhere in the UK. Despite this, the area consists of an extremely divergent population that none of the eight 2011 OAC Supergroups adequately describes. Further investigation revealed that this OA contains retirement proprieties that can only be bought by those over the age of 55. Similar to Figure 8.6, this provides an explanation into the divergent nature of the resident's characteristics, but unlike Figure 8.6 the presence of retirement only properties is not obvious from looking at only satellite imagery. In certain cases however, no justification can be found as to why an OA may be an extreme outlier. Figure 8.8 is of an OA that contains residential properties and a cemetery in Kingston upon Hull that is classed as an extreme outlier. However, satellite imagery and Google Street View reveal no obvious reasons why there may be a lack of social homogeneity. It can therefore be concluded that it would be most appropriate for the evaluation of the rationale for the homogeneity to be determined by those with local knowledge; or those able to use additional sources to perform more in depth evaluations.

Figure 8.6: 2011 OAC outlier in Plymouth

Figure 8.7: 2011 OAC outlier in Wokingham

Figure 8.8: 2011 OAC outlier in Kingston upon Hull

The standard geodemographic assignment of a cluster to an areal unit simplifies the complexity of the 2011 OAC. Analysis of the homogeneity of the 2011 OAC between the three hierarchical levels, the individual clusters and the geographic areas, explores this complexity and provides detailed information about the classification and how it was formed. The extent to which the identification of less homogenous clusters or geographic areas poses an issue for users of the 2011 OAC is dependent on how the classification is used. However, it should be expected that within any population as large as the UK's there will always be some areas on the peripheral, and with the release of metadata like SED values as an accompaniment to the 2011 OAC, these areas can be identified and further validation can be performed by users of the classification.

The ability to identify atypicality in the 2011 OAC is relatively straightforward. Statistics providing a broad overview, such as '2.27% of OAs and SAs assigned to Supergroups are considered outliers', provide a simple way of understanding the merits of the clustering solutions, and are particularly useful for comparing the different hierarchical levels.

However, when these statistics are applied at OA or SA level, it is much more difficult to explain why this atypicality exists within the classification. If a justification for the classification cannot be found, a fundamental problem with the methodology would be implied. However, the variation in the levels of homogeneity that exist within the 2011 OAC can for the most part be explained. The few instances that cannot be understood are due to a lack of local knowledge, which cannot be factored in when analysing the levels of homogeneity across the entire UK scope of the 2011 OAC.

8.5. Changes between 2001 and 2011

One of the key validations of the 2011 OAC is how it incorporates the change that has occurred in the UK over the last decade. Between 2001 and 2011 the population of the UK increased by 4.1 million, almost 7%, to 63.2 million (ONS, 2012l). This increase will have undoubtedly had an impact on how the geodemographic characteristics of the UK are represented. The latest expansion of the European Union, which began in 2004, has led to an increase in the total number of people from outside of the UK eligible to work in the country (Blanchflower and Lawton, 2008). It is estimated that of the total UK population in 2011, 2.7 million (4.27%) were born in other EU countries and 1.1 million (1.74%) of these were born in countries that joined the EU after 2004 (Vargas-Silva, 2013).

The impact of the overall increase in the UK population and the influx of those from outside the country is likely to manifest itself in two ways in the 2011 OAC: changes in the physical environment and changes in the social environment. The physical changes include the development of new houses or flat complexes to accommodate the increasing population. The social changes are caused by the influx of individuals from outside the UK coming to the country to live and work, and thereby changing the social characteristics of the location where they settle.

The changes in the physical environment can be simply quantified as they are easily identifiable. Figure 8.9 are two satellite images of an area in Basingstoke, a town in the English county of Hampshire. The image on the left is from 2001 and the image on the right is from 2011. The development of two large flat and housing complexes in the intervening years is clearly visible in the central area of the 2011 map. In 2001, the land appears to have been a mixture of light industry and open green spaces. The change in land use and subsequent increase in population is likely to have had an impact on the geodemographic output of the area. Such developments have been repeated across the UK and are not inherently different to those that took place over previous inter-censal periods. Changes between old and new geodemographic classifications are therefore not uncommon, but an important part of any new system is ensuring it incorporates the latest developments.

Figure 8.10 presents a comparison between the 2001 OAC and 2011 OAC for the same area of Basingstoke. Firstly, it is clear that the number of OAs in 2011 has expanded to reflect the increases in population in the areas with the new housing. Secondly, the 2001 OAC classified the areas where the new housing is located and the existing housing to the south of it as the same 'Typical Traits' Supergroup. This is a result of the previous lack of residents in the formerly unpopulated areas in 2001 needing to be incorporated with the pre-existing residential areas to ensure total geographical coverage of the OAs.

Figure 8 9: Basingstoke in 2001 and 2011

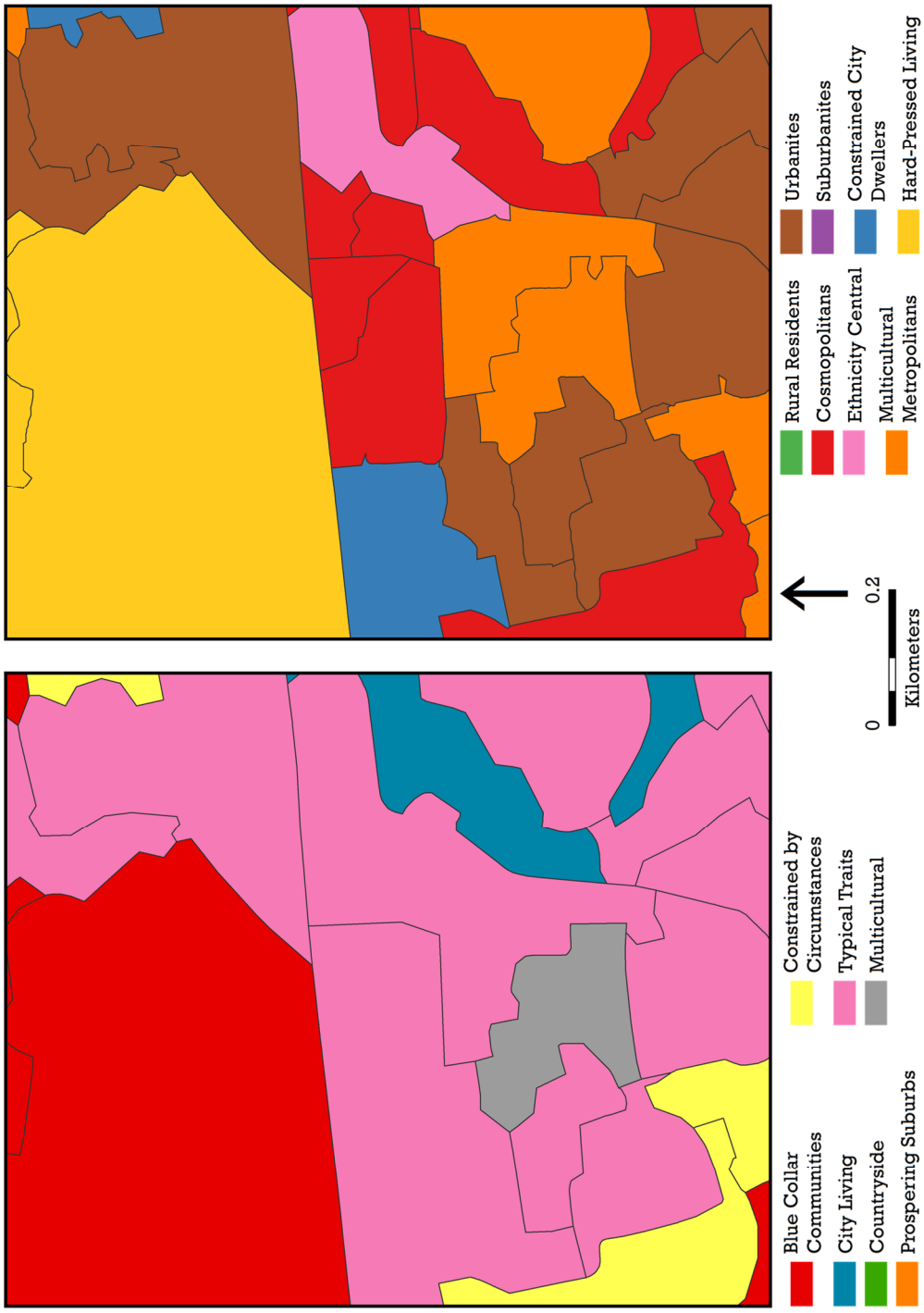


Figure 8.10: The 2001 OAC and 2011 OAC in Basingstoke

The new 2011 OA boundaries now separate the older residential areas and the newer developments into three different Supergroups. The new housing is classified as 'Cosmopolitans' and the older residential areas as 'Urbanities' (the equivalent cluster to 'Typical Traits' in the 2001 OAC) and 'Constrained City Dwellers'. This example shows that the changes in population and subsequent modifications to the OA (and SA) geography are picked up by the 2011 OAC. In the Basingstoke example, the new housing developments contain populations with sufficiently different characteristics to those who live in the more established residential areas. This results in them being assigned to different 2011 OAC Supergroups. Not all new developments that have been created in the UK since 2001 are likely to produce the same clear distinction. It is just as plausible that new developments contain populations which are not dissimilar to those in the existing surrounding housing. In such cases the 2011 OAC is likely to classify the older and newer areas as the same Supergroup. The extent to which this is true can only be assessed on a case-by-case basis, and will form part of the larger evaluation that users of the 2011 OAC will perform when assessing the potential use of the classification for their needs.

The social changes that have occurred across the UK are well documented elsewhere (such as 1.74% of the UK's population in 2011 being individuals born in countries that joined the EU after 2004), but the translation of these statistics to their impact on neighbourhoods can be harder to identify. To assess how social changes are manifested in the 2011 OAC, the city of Southampton in southern England is used as an example. Figure 8.11 shows the distribution of both the 2001 OAC Supergroups and the distribution of the 'White Other' population across the city in 2001. The correlation coefficient between the 'White Other' population and the 'not born in the UK' population is 0.85. The 'White Other' population can therefore act as proxy to identify individuals not from the UK who are white. The 'White Other' population for each OA has been calculated as a percentage of each OAs total population. The 'White Other' population made up 2.57% of Southampton's total population in 2001, with a maximum concentration of 19% been recorded in a single OA. The areas with higher concentrations of the 'White Other' population correlate to the OAs that are assigned to the 'City Living' Supergroup.

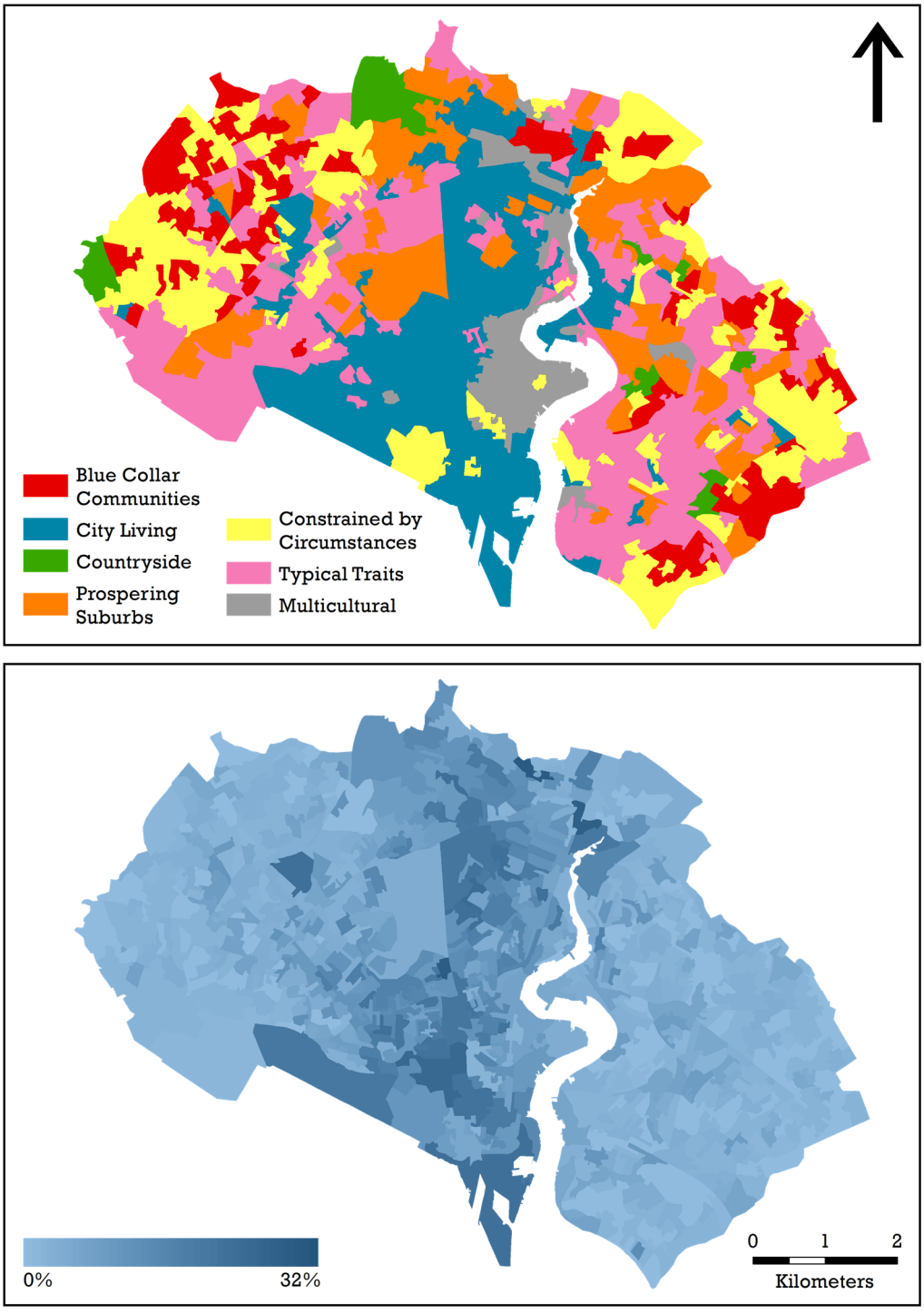


Figure 8.11: The 2001 OAC Supergroups and Southampton's 'White Other' population in 2001

The EU expansion has had a significant impact on Southampton, as shown in Figure 8.12, with a large increase in the 'White Other' population. In 2011 the 'White Other' population accounted for 7.37% of Southampton's total population, a 212% increase from 2001. This increase has resulted in the 'White Other' population becoming more geographically disbursed across Southampton. There are now three Supergroups which account for the areas with higher 'White Other' densities, 'Cosmopolitans', 'Multicultural Metropolitans' and 'Ethnicity Central', in comparison to the single Supergroup in 2001. All three of these 2011 OAC Supergroups have an above national average number of persons born in the newer EU member states.

Unlike the 2001 UK Census, more specific ethnicity outputs from the 2011 UK Census allows for analysis of subsections of the 'White Other' population. Figure 8.13 shows the distribution of the Polish population in Southampton, and the Groups from the 'Multicultural Metropolitans' and 'Cosmopolitans' 2011 OAC Supergroups. The 'White Polish' population accounted for 3.22% of Southampton's total population in 2011, and is found in two concentrated geographic locations in the city. These areas correspond to the 'Rented Family Living' Group to the west and the 'Comfortable Cosmopolitans' and student oriented Groups in the central areas of Southampton.

No 'Polish' variable existed as a category in the 2001 UK Census outputs, but even if it had, it is unlikely the high concentrations in the white Polish population found in 2011 in the western areas of Southampton would have existed in 2001. As shown by the Figure 8.11, in 2001 that area of the city had much lower concentrations of the 'White Other' population. The 2001 OAC classifies these areas as predominately 'Typical Traits' and to a lesser extent a mixture of the 'City Living' and 'Constrained by Circumstances' Supergroups. The 'Typical Traits' and 'Constrained by Circumstances' both have below the national average of persons born outside of the UK. As such, it is unlikely that any area assigned to these Supergroups would have had a high proportion of the 'White Other' population.

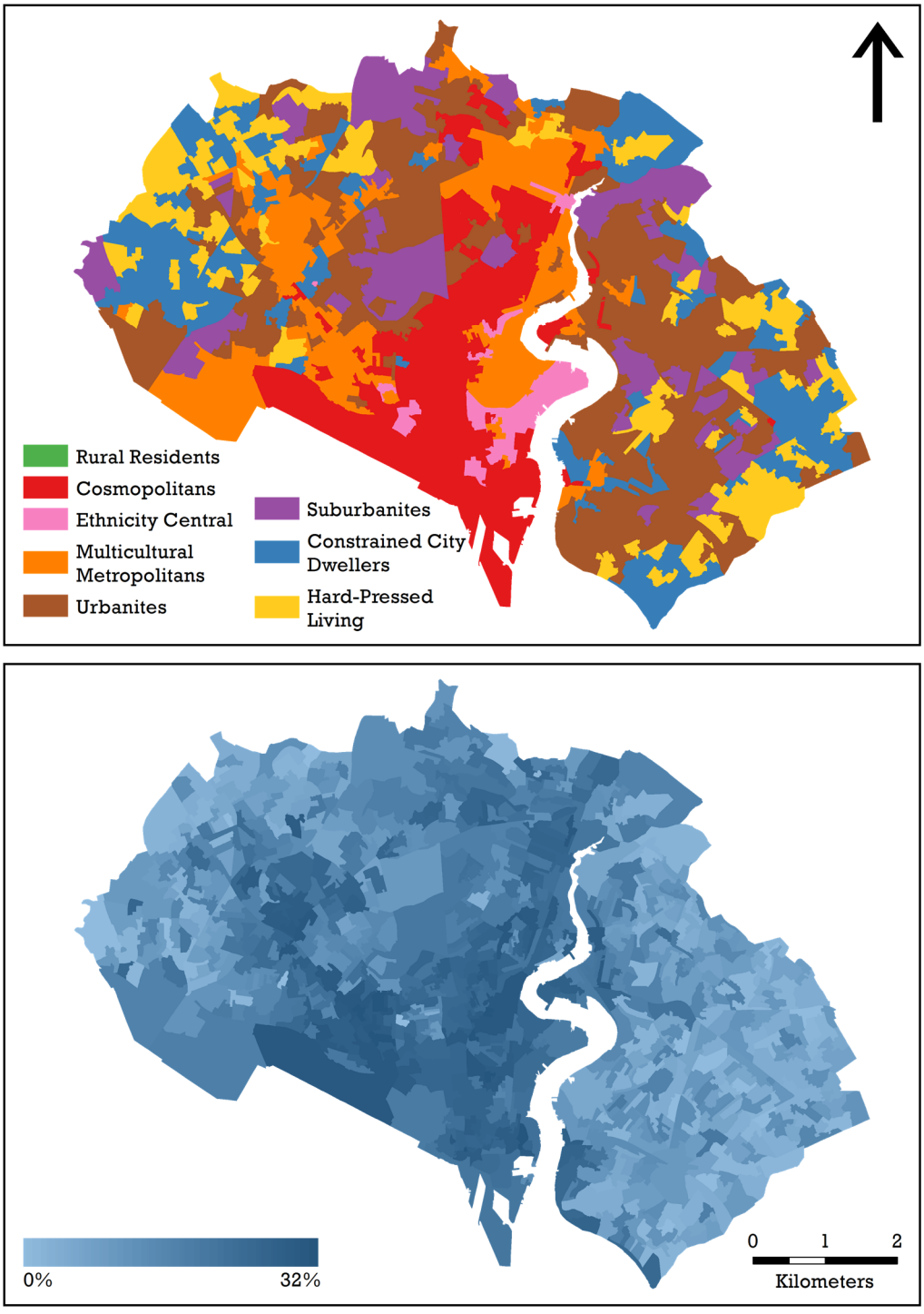


Figure 8.12: The 2011 OAC Supergroups and Southampton's 'White Other' population in 2011

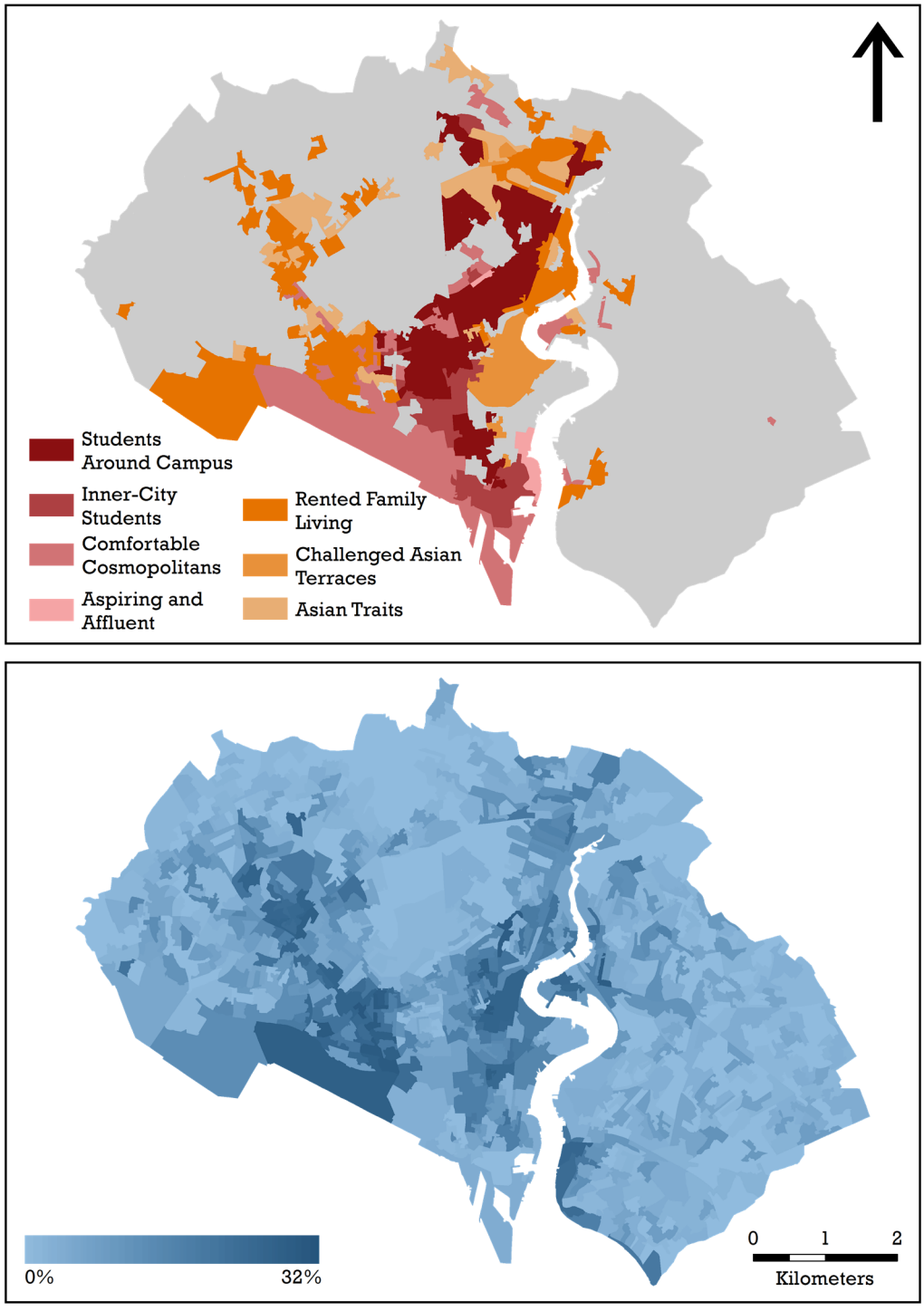


Figure 8.13: The 2011 OAC Groups and Southampton's 'White Polish' population in 2011

It would appear that since 2001, and most likely since 2004, the increase in the 'White Other' population has resulted in certain areas in Southampton becoming more multicultural. This fact can be evidenced by the assignment of certain OAs to the 2011 OAC 'Multicultural Metropolitans' Supergroup to areas that previously had a below average ethnic mix. As these assignments correlate with areas that now have a higher concentration of 'White Other' and in particular, 'White Polish' populations, is an indication that the 2011 OAC has identified social changes in those areas.

The Southampton example provides evidence that the changing social structures of neighbourhoods in the UK since 2001 can be identified with the 2011 OAC. Similar to the Basingstoke example however, it only provides evidence for one comparatively small geographical area in the UK. As such, it does not guarantee that all social change over the past decade will be identified.

The suggestion that the Southampton example is therefore not reflective of how the 2011 OAC performs with social change across the rest of the UK is difficult to assess without wider validation exercises. Similar to the physical changes in the UK, users of the 2011 OAC will be best placed to perform more detailed evaluations, based on their local knowledge of the accuracy of the 2011 OAC. It can however be said that based on the examples from Basingstoke and Southampton, the 2011 OAC appears to be adept at incorporating the change in the physical and social environment of the UK that have occurred since 2001.

8.6. Ground-truthing

An important validation exercise for the 2011 OAC was looking at *how* the different clusters represent each area. During the construction of the 2011 OAC, great significance was placed on ensuring that the final clusters were both as evenly distributed as possible and sufficiently different from one and other. Although these aims were judged successful, both from a statistical perspective and because the final output looked visually representative (Mandelbrot, 1982b), the ability to assess quantitatively whether the final clusters were archetypal of an area was more difficult. Therefore, a qualitative evaluation of areas assigned to different clusters was required to assess whether the cluster names and descriptions adequately summarise the characteristics of neighbourhoods. This 'ground-truthing' of a classification to analyse whether the data corresponds to the reality of an area has been undertaken in previous studies. Vickers

and Rees (2011) detail the more expansive measures undertaken with the 2001 OAC once it had been released to a wide user base. At the time of writing, the 2011 OAC has yet to be released to a wide user base, so the ground-truthing performed so far is comparably limited in scale.

To 'ground-truth' the 2011 OAC, 129 first year Geography undergraduate students at University College London (UCL) were enlisted. Each student was given the names (Table 7.10) and pen portraits for each Supergroup (see Section C.1 in Appendix C). They then visited up to ten unique postcode sectors within Greater London and were asked to record the following information:

- The assigned 2011 OAC Supergroup for that area.
- Does the assigned Supergroup best describe this area out of the 8 Supergroup Options?
- If the answer to the previous question was 'No', what alternative Supergroup best describes this area?

These questions were designed to elicit a response that could be quantified (i.e. 'yes' or 'no') so that the results could be analysed to provide an understanding of the effectiveness of the 2011 OAC. However, it is recognised that there are known limitations in using direct observations for data collection; participants have access to different bodies of knowledge (DeWalt and DeWalt, 2002) and human observations are inherently biased (Kawulich, 2005). These limitations create an inconsistency in the results. Dixon and Leach (1978) commented that increased numbers of responses creates more meaningful conclusions with Gould and White (1974) providing a specific example of using students to bolster response rates. The 129 participants cannot however be considered free from group or individual bias. As the scope of the exercise was limited to postcode sectors being visited by one individual only, inconsistencies in the interpretation between areas are likely to be present in the results. It is therefore important to understand the limitations of the ground-truthing exercise when considering the findings.

In total, the students visited 974 unique postcode units. The results for whether they believed the assigned Supergroup best described each area are shown in Table 8.6 and Figure 8.14. No students visited any area assigned to 'Rural Residents' and the results for the 'Constrained City Dwellers' and 'Hard-Pressed Living' Supergroups are discounted

because of the small sample size. The results from the remaining five Supergroups show that almost half of the postcode units visited were assigned to 'Cosmopolitans', and over a quarter were in the 'Ethnicity Central' Supergroup. Although there is a disparity between the Supergroups visited, the results allow the suitability of the cluster assignments to be evaluated.

Across the areas visited, the students agreed more frequently than they disagreed that the assigned Supergroup offered the best representation of an area. On average for every three postcode units visited, two would have the most appropriate Supergroup assignment and one would not (a ratio of 2:1). These values remained consistent between the Supergroups, with only the 'Suburbanites' Supergroup showing significantly better results (a ratio of 7:1). However, this is likely to be due to the smaller number of areas visited that were assigned to that Supergroup, rather than because it is fundamentally better at describing an area.

Table 8.6: Ground-truthing the 2011 OAC: Does the Supergroup best describe the assigned area out of the 8 Supergroup options?

Supergroup	**Yes**	**No**	**Proportion of responses**
Rural Residents	0% (0)	0% (0)	0% (0)
Cosmopolitans	65% (288)	35% (153)	45.3% (441)
Ethnicity Central	68% (191)	32% (90)	28.9% (281)
Multicultural Metropolitans	65% (110)	35% (59)	17.4% (169)
Urbanites	67% (43)	33% (21)	6.6% (64)
Suburbanites	88% (14)	13% (2)	1.6% (16)
Constrained City Dwellers	100% (1)	0% (0)	0.1% (1)
Hard-Pressed Living	100% (2)	0% (0)	0.2% (2)

(Counts are in brackets)

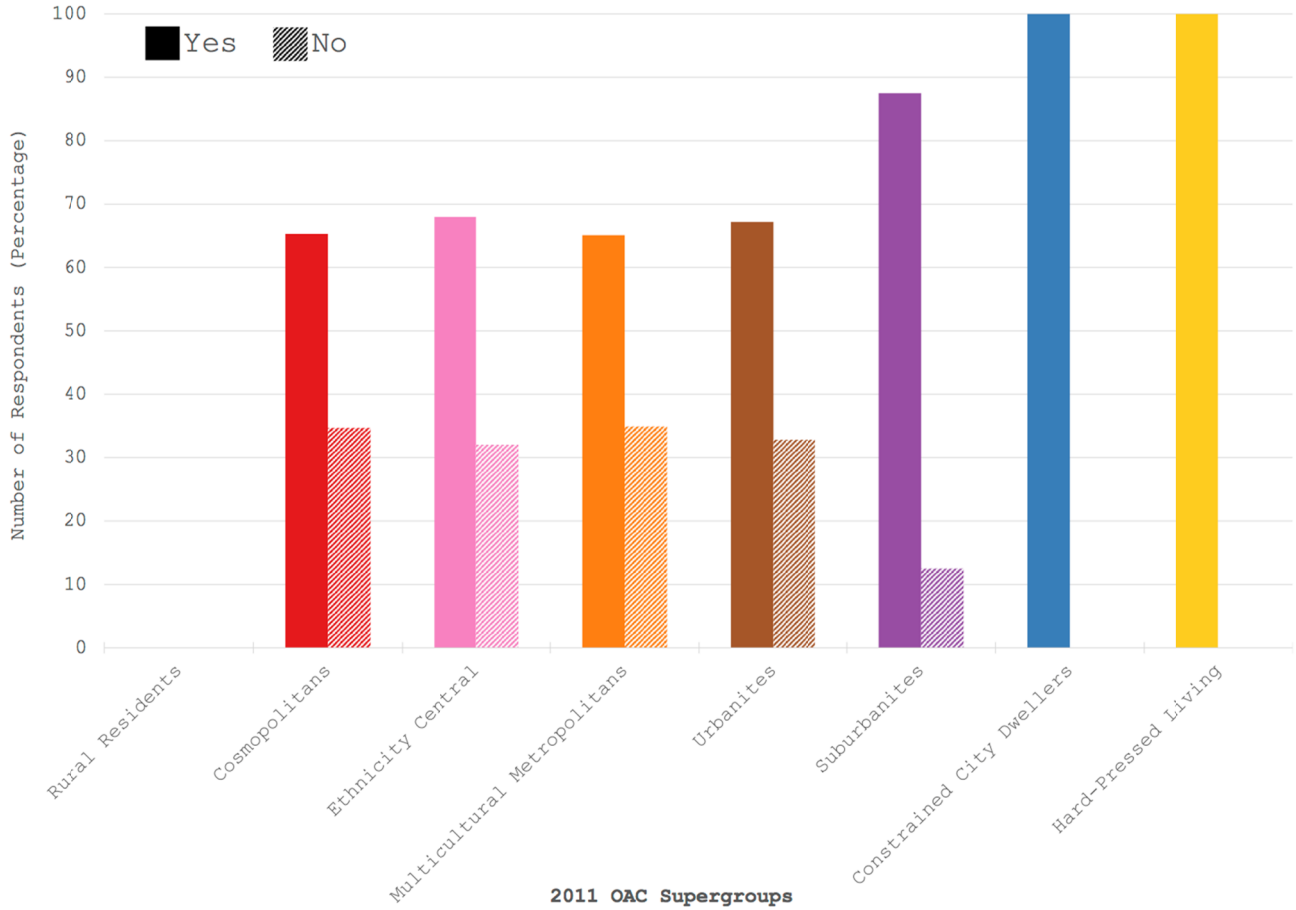


Figure 8.14: Responses to if the 2011 OAC Supergroup assignments are the best option for an area

Figure 8.15 illustrates the spatial distribution of these results across London. It is notable that there is no significant clustering of results visible, with perhaps the exception of participants who appeared more likely to agree with the Supergroup assignment in central areas. This lack of any notable clustering is a positive outcome for the 2011 OAC. A fundamental principle of geodemographics is that a clustering output is correct frequently enough to justify its existence. The fact that the areas identified by students as incorrectly classified were distributed across London, with no spatial clustering, indicates that there is not a systematic problem with the designation of areas to the various 2011 OAC clusters.

The results from the 129 participants indicate the 2011 OAC most commonly assigns the best Supergroup option to an area, suggesting from a subjective perspective the classification provides the best representation possible more often than not. Although the nature of the ground-truthing exercise causes the results to be exposed to elements of subjectivity, this is at least partially counteracted by the benefits of having a large cohort taking part in the exercise.

The results from this ground-truthing exercise are positive for the 2011 OAC, the focus of the exercise solely on London makes it difficult to predict how representative the results are for the rest of the UK. Additionally, it is difficult to determine the appropriateness of the classification at the Group and Subgroup levels. As clusters become more specific at these levels, observations by participants of the physical environment are increasingly less likely to identify differences between areas, namely those that cannot be observed directly. A more in-depth ground-truthing process would be required to address this, such as that performed by Vickers and Rees (2011).

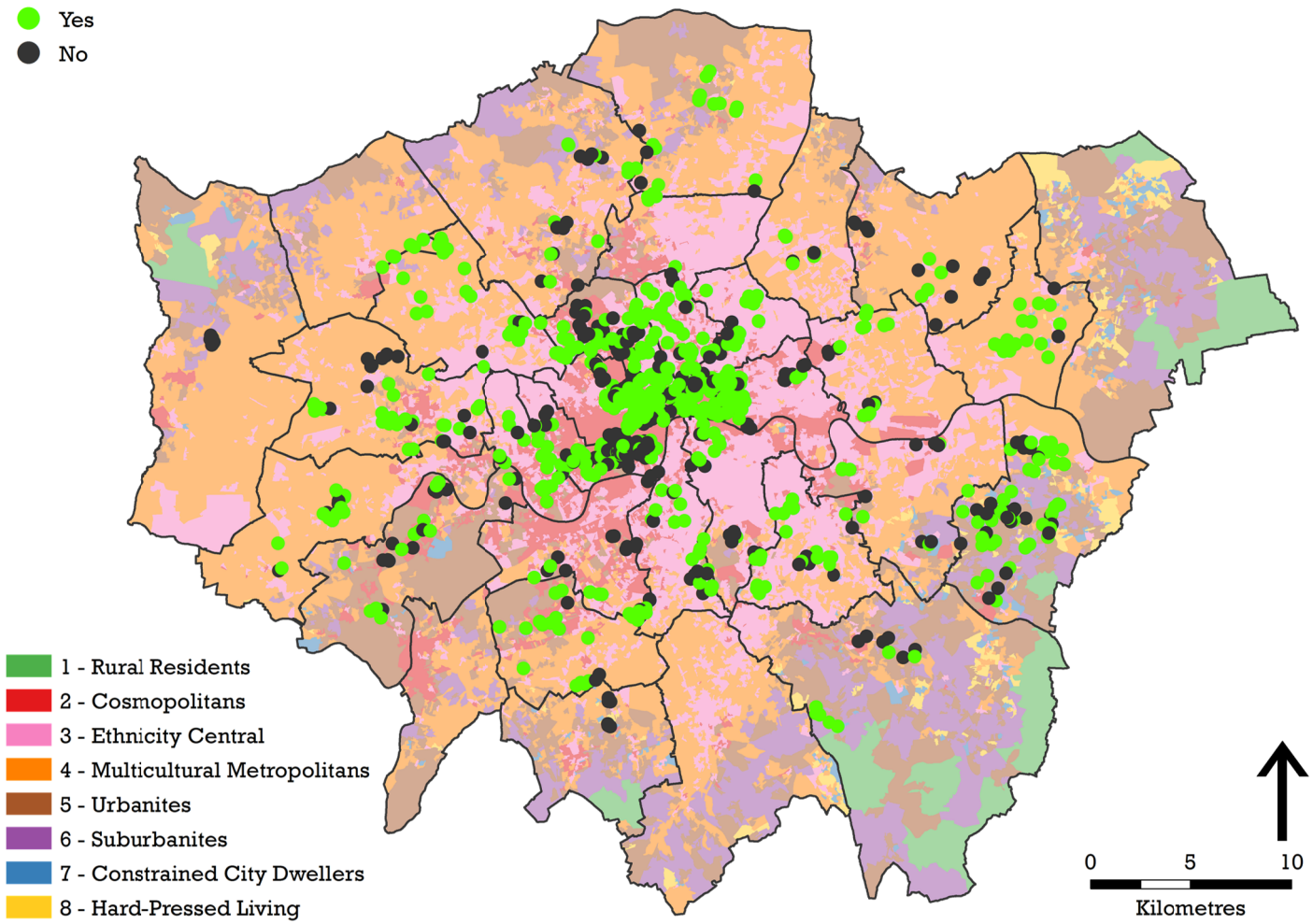


Figure 8.15: The locations in London where the 2011 OAC Supergroups assignment is considered the best or not the best option

The 33% of postcode units that were judged as having been assigned to an inappropriate Supergroup were explored to ascertain what, if any, reasoning existed to explain their occurrences. Table 8.7 is a confusion matrix that identifies an alternative Supergroup assignment where the original was felt inappropriate. In the opinions of the participants, the Supergroup most commonly found to be incorrectly assigned was 'Cosmopolitans'. In 71% of these cases, the participants felt that the 'Urbanites' Supergroup would have been more appropriate. This can also be seen on Figure 8.16, where the spatial locations of the alternative Supergroup assignments have been mapped. There is a small cluster of postcodes in the central areas of London where participants recommended that 'Urbanites' would have been a more appropriate Supergroup, but aside from these there are no other large visible concentrations of alternative Supergroup suggestions.

Table 8.7: Ground-truthing the 2011 OAC: If the assigned Supergroup does not best describe the assigned area, which alternative Supergroup does?

		Assigned Supergroup							
	Supergroup	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Suggested Supergroup	**1**		2% (3)	1% (1)	0% (0)	10% (2)	0% (0)	0% (0)	0% (0)
	2	0% (0)		21% (20)	5% (3)	10% (2)	0% (0)	0% (0)	0% (0)
	3	0% (0)	9% (14)		22% (13)	0% (0)	0% (0)	0% (0)	0% (0)
	4	0% (0)	4% (6)	30% (28)		0% (0)	0% (0)	0% (0)	0% (0)
	5	0% (0)	71% (108)	28% (26)	22% (13)		0% (0)	0% (0)	0% (0)
	6	0% (0)	2% (3)	3% (3)	22% (13)	40% (8)		0% (0)	0% (0)
	7	0% (0)	12% (18)	13% (12)	2% (1)	5% (1)	33% (1)		0% (0)
	8	0% (0)	0% (0)	4% (4)	26% (15)	35% (7)	67% (2)	0% (0)	

(Counts are in brackets)

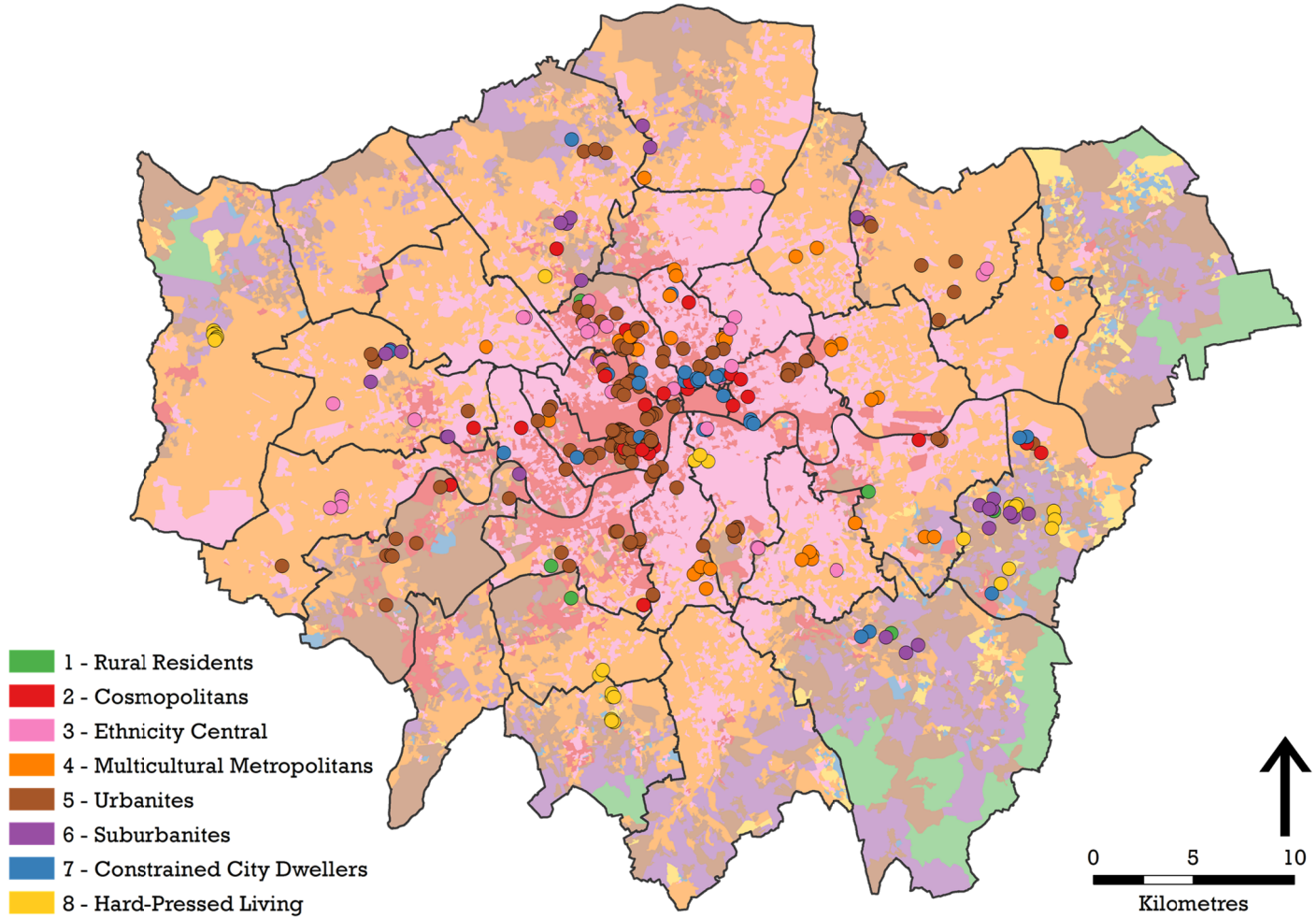


Figure 8.16: The locations in London where an alternative 2011 OAC Supergroup assignment has been suggested

In relation to the wider 2011 OAC ground-truthing exercise these results could have been dismissed. They only represented the minority view of the participants and because of unquantifiable factors such as the misinterpretation of the names and pen portraits by the participants. It was therefore necessary to undertake analytical analysis to ascertain the likelihood that the areas identified as being assigned to the wrong Supergroup contained characteristics which would imply this was the case. This analysis was based on the SED values for the OAs in the areas visited, and their respective Supergroup assignment. Table 8.8 details the outcome of this analysis where a comparison is made between the SED values per Supergroup with the 'correct' and 'incorrect' Supergroup assignments. The two values for each Supergroup provide a summary measure to indicate the level of cluster homogeneity. Larger SED values indicate that the level of homogeneity in that Supergroup is lower, increasing the possibility that characteristics from other Supergroups are more prevalent (as discussed in Section 8.3). A supporting argument could therefore be made for the 'incorrect' Supergroup assignment if their SED values were higher than those assigned 'correctly'.

As shown in Table 8.8 however, there are no notable differences between the SED values for 'Ethnicity Central' and 'Cosmopolitans'. The areas judged to be incorrectly assigned to these Supergroups are actually more homogenous. Figure 8.17 maps the SED values, and show that areas which have been judged as having the wrong Supergroup assignment are in areas that can be considered to be more uncertain. It is however the case that areas considered to have the correct Supergroup assignment can also be found in similarly uncertain areas. The mean SED values offer limited opportunity for inference as only a summary measure is provided to indicate the overall level of homogeneity in the target areas. The small difference in values would indicate that there is no evidence to suggest the areas identified by participants as being incorrectly assigned to a Supergroup are any more likely to be incorrect than the areas identified as being 'correct'.

Table 8.8: The mean SED values for the ground-truthed 2011 OAC Supergroups

Supergroup	Best Supergroup assignment	Not the best Supergroup assignment
Rural Residents	N/A	N/A
Cosmopolitans	1.106	1.082
Ethnicity Central	0.928	0.910
Multicultural Metropolitans	0.931	0.934
Urbanites	0.963	0.959
Suburbanites	1.014	0.987
Constrained City Dwellers	1.048	N/A
Hard-Pressed Living	0.858	N/A

Based on the premise discussed in Section 8.3 that each OA and SA in the UK contains some element of all 8 Supergroups, further analysis was carried out to quantify the likelihood of an area conforming to an alternative Supergroup assignment. As previously discussed, Slingsby et al. (2011) uses examples from the 2001 OAC to show particular OAs which could have been assigned a selection of Supergroups, as the characteristics of those areas shared multiple traits with multiple clusters. These conclusions were not based on the mean SED value of a Supergroup, but on how close the mean SED value for the assigned Supergroup is to the next best option (defined as having the second lowest SED value).

Table 8.9 presents the mean SED difference between the assigned Supergroup and the second best option. The smaller the value, the more in common the area has to the second Supergroup option. As such, an 'incorrect' assignment to an area would have a lower value than areas designated as 'correct'. This would imply that the area has characteristics of more than one cluster, and is therefore less likely to conform to the assigned Supergroup archetype. The results in Table 8.9 are inconsistent. For example, the areas judged to the assigned to 'Cosmopolitans' incorrectly have a 19% smaller SED value, yet for 'Urbanites' it is 27% larger. This lack of consistency between Supergroups and the relatively small differences between 'correct' and 'incorrect' assignments make it difficult to draw any decisive conclusions. It does however appear unlikely that SED values can be used to explain the rationale used by participants in judging the suitability of a Supergroup to the areas they visited.

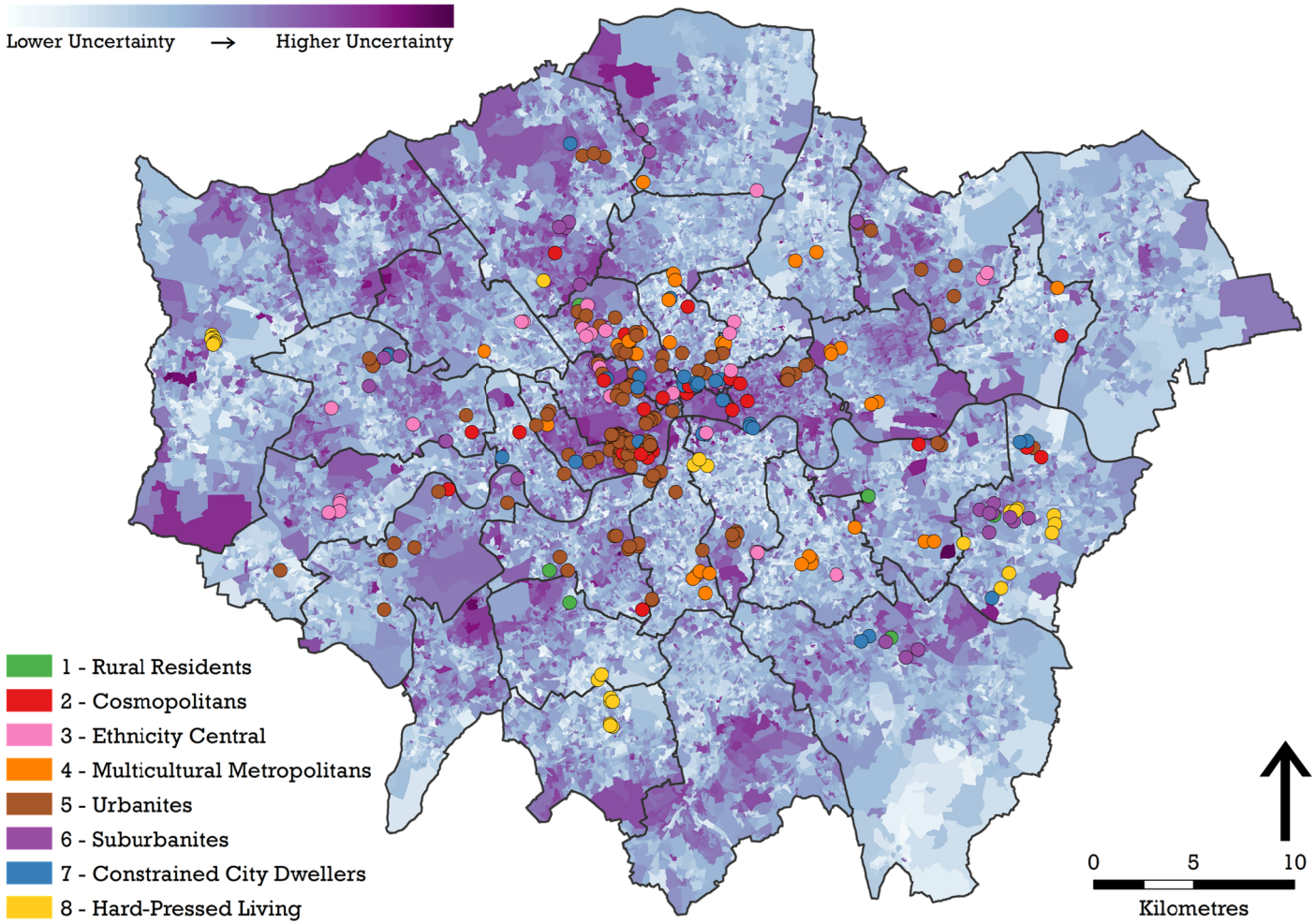


Figure 8.17: The locations in London where an alternative 2011 OAC Supergroup assignment has been suggested overlaid on an uncertainty

Table 8.9: The mean SED values between the assigned ground-truthed 2011 OAC Supergroup and the next 2011 OAC Supergroup

Supergroup	Best Supergroup assignment	Not the best Supergroup assignment
Rural Residents	N/A	N/A
Cosmopolitans	0.267	0.218
Ethnicity Central	0.226	0.185
Multicultural Metropolitans	0.246	0.223
Urbanites	0.123	0.157
Suburbanites	0.113	0.075
Constrained City Dwellers	0.115	N/A
Hard-Pressed Living	0.177	N/A

This ground-truthing exercise suggests that the 2011 OAC frequently, at least at the Supergroup level, provides the best representation possible. The analysis of areas identified as having the wrong Supergroup assignment suggests that there is no statistical reason for this. The conforming of an area to a Supergroup was due to the way in which each participant in the ground-truthing exercise interpreted the area they were in.

The limited scope of the ground-truthing, both in geographic scale and the questions asked does have an impact on the validity of the findings. For instance, the respondents were asked whether the assigned Supergroup to an area was the best out of the eight options, not whether it provided an accurate summary of that location. To truly judge how reflective the cluster assignments of the 2011 OAC are, a more expansive, in terms of geographic coverage, and thorough, in terms assessing the different hierarchies of the classification, ground-truthing exercise would be required. This could be either achieved in the same way as Vickers and Rees (2011) or be performed by individual users when assessing the appropriateness of the 2011 OAC for their needs.

8.7. Conclusion

The analysis of how the 2011 OAC performed across five different validation categories indicates that the classification is robust. As with other geodemographic classifications, it will provide a subjective representation for the entirety of its geographic extent, and can be considered reliable enough across the majority of the UK for it to be a useful tool. The 2011 OAC is an open geodemographic classification which allows for analysis of all outputs of the clustering process. Subsequently, cluster assignments which may not offer the best representation of population characteristics in an area, as may be indicated by SED values, can be explored to help understand why this may be the case.

The validation of the 2011 OAC has demonstrated that the selection of variables which provided the best coverage of the UK's population characteristics has led to the inclusion of some statistically sub-optimum variables. Although these variables can be clearly identified through the Lorenz curves and Gini Coefficients, it can be concluded that their inclusion does not adversely impact the 2011 OAC. Although the inclusion of variables which make the formation of homogenous clusters more difficult is not advisable in any geodemographic classification, the interaction of the 60 variables of the 2011 OAC to form the distinct clusters which structure the classifications hierarchy indicates that the chosen variables have performed well. The utilisation of both quantitative and qualitative assessments to fully understand the performance of a geodemographic classification is essential. A consideration of both components has allowed the 2011 OAC to contain 110 unique clusters.

Although the qualitative aspect of variable selection aided the creation of distinct clusters, the sheer number of OAs and SAs which cover the UK meant that a quantitative focus was required to assess the certainty of cluster assignments to each area. The use of the SED values as a dissimilarity measure for the k-means clustering algorithm allowed for the evaluation of the level of homogeneity each OA or SA had to its assigned cluster. This analysis revealed that OAs in Scotland had a higher propensity to differ from the average characteristics of their assigned cluster when compared to other parts of the UK. This was especially evident for the 'Rural Residents' Supergroup. Consequently, residents who lived in areas classified as this cluster in Scotland were likely to have characteristics more divergent from the cluster average than elsewhere in the UK.

The atypical nature of Scotland compared to the rest of the UK can be attributed to the design of the Census geography rather than any natural propensity for the population to

have divergent characteristics. The design of the OAs in Scotland to have smaller population and household threshold values act to increase the number of atypical areas found in Scotland. This impacts the 2011 OAC, with analysis of the homogeneity of Supergroups between different English Regions, Wales, Northern Ireland and Scotland, provided more evidence to Scotland's atypical characteristics. This analysis highlighted areas classified as 'Constrained City Dwellers' and 'Cosmopolitans' as the least homogenous. The less homogenous nature of 'Constrained City Dwellers' can be explained by the prevalence of the Supergroup in Scotland. The large number of less homogenous areas outside of Scotland assigned to the 'Cosmopolitans' Supergroup suggest that there are a wider range of population characteristics which are identified and incorporated into this cluster. Identifying the reasons for this can be difficult, especially when evaluating individual OAs or SAs.

Analysis of the homogeneity of the cluster assignment across the whole of the 2011 OAC revealed that SED values provided a reliable method of identifying the OAs or SAs which were least likely to conform to their assigned cluster. Although an area can be easily identified as likely to be less homogenous, explaining why this is the case poses complications. The utilisation of local knowledge to explain these occurrences would be required.

Although the assessment of cluster assignment and homogeneity was a key component in the validation of the 2011 OAC, it was also important to ensure that the classification was able to identify change in the UK since 2001. The example of new housing built in Basingstoke and the increase in the 'White Other' population in Southampton, provided opportunities for comparison of how the 2011 OAC classified these areas with the 2001 OAC. In both instances, the 2011 OAC clusters accounted for the changes. In Basingstoke, the new housing was classified into a different Supergroup than the surrounding older housing, which was a reflection of the differing population characteristics present in the areas.

In Southampton, the social change caused by the increase in the 'White Other' population can be seen in the 2011 OAC. Areas classified by the 2001 OAC as having a low number of people born outside the UK, are now classified into clusters which have an above average number of 'White Other'. This variable has a significant positive correlation with people born outside the UK. The two examples from Basingstoke and Southampton do not guarantee that all change that has occurred since 2001 will be incorporated into the

2011 OAC, but it does suggest that the classification has the ability to incorporate the physical and social environmental change that has taken place.

The final category of validation performed on the 2011 OAC was the limited ground-truthing exercise performed by students from UCL. Although the results have to be tempered by the inherent problems that exist when conducting an exercise of this nature, the results are positive for the 2011 OAC. A greater number of students agreed that the Supergroup assignment to an area was the best option available. This would suggest that assignment of a 2011 OAC cluster to an OA or SA will be correct more often than it is wrong. Although the limited geographical coverage and the assessment of only the Supergroup level of the 2011 OAC, this does, however, mean that the applicability of the results to all hierarchical levels of the classification and to areas outside of London is unknown.

The validation exercises performed on the 2011 OAC suggest that it is a robust classification that consists of distinct clusters, which incorporate change in the UK since 2001 and is correct more often than it is wrong. The inclusion of measures such as utilisation of SED values to quantify variances in cluster assignment are designed to provide a useful way for users to explore the underlying complexity of the 2011 OAC. The extent to which these measures are used, and the ultimate usability and robustness of the 2011 OAC will however be determined by users, and whether they believe the classification to be reliable enough to be used for their intended purposes.

Chapter 9

Conclusions and Future Work

9.1. Introduction

This chapter summarises the research undertaken to create the 2011 Area Classification for Output Areas (2011 OAC). The success of the primary aim of the project to create a new open geodemographic classification of the UK is assessed in Section 9.2. The five secondary aims which were developed to provide focus for the construction of the 2011 OAC are individually evaluated. The extent to which these secondary aims were fulfilled and their contribution to the research aims of the project are assessed. Although the creation of the 2011 OAC is an important milestone, there is a wide array of further development that could be undertaken for future use. The classification has been designed with a range of different users in mind, such as local government, academia and commercial companies. All of these are likely to use the 2011 OAC for different tasks. Some of these potential applications are discussed in Section 9.3 and how elements of the 2011 OAC, such as the use of open source software and transparent methodology, allow the classification to be adaptable to the geodemographic requirements of users.

Section 9.4 discusses the lifespan of the 2011 OAC and how the future of the classification will be determined by continuing research into open geodemographics. This future work will affect whether the 2011 OAC can be updated in a similar fashion to current commercial systems, or whether the creation of temporal uncertainty measures offers the best solution based on current data availability in the UK. Finally, Section 9.5 draws together the main research topics of the project. The unique status of the 2011 OAC in the UK geodemographic classification marketplace is examined in addition to its future and contribution to open geodemographics research.

9.2. Summary of research aims

The main aim of the project was to take advantage of the release of 2011 UK Census data to create a new open geodemographic classification of the UK. Creating any

geodemographic classification is a complex task, with the combination of multiple methodological aspects to form the final outcome. Although this overall aim has been achieved, simply creating a classification is insufficient to declare the project a success. The extent to which the 2011 Area Classification for Output Areas (2011 OAC) can be regarded a success is dependent on the outcomes of the secondary aims which provided the necessary focus to create the new classification. These secondary aims are re-stated below and the extent to which they have been fulfilled is assessed.

i) To create a new open, transparent and reproducible methodology.

This aim was primarily focused on improving the methodology used to create the 2001 OAC. This is discussed in Chapter 6, and its implementation is detailed in Chapter 7. The methodology of the 2011 OAC can be considered as a more robust version than that of the 2001 OAC. The need to create a completely new methodology to that used to build the 2001 OAC was deemed unnecessary as it was concluded that this would repeat research previously explored by Vickers (2006). The project instead focussed on improving certain methodological aspects used in the creation of the 2001 OAC.

The two distinguishing features of the 2011 OAC methodology are the use of open source software to perform the majority of the statistical and clustering operations, and the testing of multiple data preparation techniques. The use of the R program (R Development Core Team, 2011) and the subsequent release of the code makes the 2011 OAC methodology completely open and transparent. Unlike the 2001 OAC, which used SPSS, this allows universal access to the code free from any licensing restrictions. The code can then be utilised to either recreate the 2011 OAC, or to adapt and apply it in different scenarios to aid in the creation of bespoke geodemographic classifications.

The testing of multiple data preparation techniques increased the likelihood that the 2011 OAC would provide an optimum representation of the UK's population. The techniques used to select the variables for the 2001 OAC and 2011 OAC did not vary significantly. However, the 2011 OAC tested multiple methods of rate calculation and data transformation, rather than relying solely on the percentages and log methods utilised in the 2001 OAC. Section 6.5 discusses these different methods. This testing produced multiple datasets, the best of which was selected on the basis of that which produced the optimum clusters. This decision was based on what looked the most right (Mandelbrot, 1982b), rather than the fulfilment of any statistical criteria.

ii) To consult with users to determine what their requirements are for the classification.

This aim was focussed on identifying the requirements of potential users of the 2011 OAC. Chapter 4 details the processes and underlying theory behind the '2011 OAC user engagement'. The process explored two key concepts. Firstly, the need to solicit the views of potential users to ascertain their priorities for the 2011 OAC. Secondly, the extent to which the 2011 OAC could be constructed to reflect the general consensus of potential users, thereby being the most use to the greatest number of people.

A total of 38 responses to the user engagement were received. A general consensus was not found on many issues, but there were some general themes that could be identified. These themes were summarised as six points, three of which had already been identified from the literature and utilised in the secondary aims: 'the need to evaluate the effectiveness of the 2011 OAC'; 'provide additional information about the outputs of the 2011 OAC'; and 'Open Data to have a role with the 2011 OAC'. Of the other three themes, only 'using the best possible data source(s) for the 2011 OAC' required a decision to be made. There were two choices available in the selection of data source(s). The first was to create the 2011 OAC at the finest granular level and utilise solely the 2011 UK Census, which could not be updated easily. The second was to incorporate Open Data sources, giving the 2011 OAC the potential to receive regular updates, although this would only be available at coarser levels of geography. It was concluded that the importance of having the classification available at the smallest levels of geography was more important than the potential for updates. Additionally, the breadth of coverage, detail and accuracy offered by the UK Census cannot be currently replicated with other Open Data sources. The 2011 OAC was therefore ultimately constructed from only 2011 UK Census data.

The final two points, 'the 2011 OAC to be a general purpose geodemographic classification' and 'the need to publicise the 2011 OAC' were more straightforward to achieve. The choice of variables from the 2011 UK Census determined the type of classification. To guarantee a general purpose classification, variables were chosen to reflect the general socio-economic characteristics of the UK's population. Section 6.6 explains the methods that were used to select variables and Section 7.2 details how these methods were used to finalise the 60 variables used to construct the 2011 OAC. Promotion of the classification by the Office for National Statistics (ONS) and via established user forums, such as the OAC User Group, will raise user awareness of the

release of the classification. It is also envisaged that a number of talks and workshops will be organised to introduce the classification to users following its release.

The 2011 OAC cannot be considered to be a consultation led geodemographic classification. However, important decisions including the data sources used and the focus of the classification were derived from user feedback. The result of this should be a classification which fulfils the requirements of the majority of users.

iii) To develop visual and descriptive outputs to facilitate users' understanding of the results produced by the classification.

This aim focused on developing outputs from the 2011 OAC that would aid user understanding of the classification. A selection of proposed outputs were included as part of the 2011 OAC user engagement, discussed in Chapter 4, with a selection of the requested outputs detailed in Section 7.4. A number of outputs produced are the same as those created for the 2001 OAC, such as radial plots. Radial plots and the bar graphs created for the 2011 OAC provide a visual summary of the characteristics of each cluster. These form the basis on which cluster names and pen portraits, or descriptions of the clusters, are based. The 2001 OAC named the Supergroup and Group levels of the classifications hierarchy, but not the Subgroups. A key output of the 2011 OAC user engagement was the desire to name the Subgroups of the new classification. This was completed after the names of the 2011 OAC Supergroups and Groups had been confirmed by the ONS. As such, all 110 clusters of the 2011 OAC have been given appropriate names to help better understand each clusters characteristics.

These descriptive outputs are complimented by visual outputs of the classification. Standard choropleth maps have been produced, displaying the geographic distribution of the 2011 OAC clusters. Section 7.4.4 discusses the issues with this type of representation, namely that it does not distinguish between densely populated urban areas and less densely populated rural areas. The use of equalising density cartograms and building maps as alternative visualisation methods of the 2011 OAC was therefore explored. Cartograms changed the size and shape of the OAs and SAs to be representative of their resident population in 2011, while the building maps only visualised the buildings in each OA or SA. Both methods provided alternative representations of the 2011 OAC. The cartograms offered a more realistic indication of the major population centres in the UK, although at the expense of geographic accuracy. Conversely, the building maps provided an accurate geographical representation of the 2011 OAC, but

could lead to misinterpretation of the classification as being accurate to the individual building level.

The different static visualisation methods used with the 2011 OAC each have their advantages and disadvantages. It will be a decision for each user as to which they prefer. The other available visualisation option is web mapping. Following the official release of the 2011 OAC the www.opengeodemographics.com website will provide a link to an online interactive map of the classification. This will overlay the classification on an OpenStreetMap (OSM) background and allows users greater freedom to explore the classification. Additionally, comma-separated values (CSV) files and ESRI shapefiles will also be released. These allow users the opportunity to import the 2011 OAC into their own geographic information system (GIS), thereby allowing them to incorporate the classification into their own workflows.

Another key output was the use of squared Euclidean distance (SED) values to assess the propensity for each OA and SA to belong to the other Supergroups of the 2011 OAC. Discussed in Section 8.3, the use of SED values allows the fuzzy characteristics of the 2011 OAC to be identified. These outputs provide a more informative summary of the 2011 OAC's Supergroup assignment, giving users an alternative way of utilising the classification.

The range of outputs for the 2011 OAC have addressed the requirements of users that were highlighted in Chapter 4 and have effectively utilised advances in GIS made since the release of the 2001 OAC. They have also expanded on the outputs that were available from the 2001 OAC, thereby allowing users greater opportunity to engage with the classification. Although the range of outputs is not extensive, the range of 'core' outputs provides sufficient summary information for the 2011 OAC to satisfy the needs of the majority of users. Furthermore, the 2011 OAC's open methodology allows users to create their own additional outputs as required.

iv) To validate the classification once complete to assess the final outcome.

This aim focused on validating the 2011 OAC to assess how representative the classification is at identifying the socio-demographic variances of the UK's population. Chapter 8 details the different validation categories used: variable specification; cluster assignment certainty; homogeneity; changes between 2001 and 2011 and ground-truthing.

The assessment of the 2011 OAC revealed that the selection of variables that provided the best coverage of the UK's population characteristics led to the inclusion of some considered statistically sub-optimum. These variables can however be identified through the use of Lorenz curves and Gini Coefficients, and it was concluded that their inclusion did not adversely impact the 2011 OAC.

Analysis of the cluster assignment certainty revealed Scotland to be distinct from the rest of the UK. The OAs in Scotland had a higher propensity to differ from the average characteristics of their assigned cluster when compared to other parts of the UK. This atypical nature of Scotland compared to the rest of the UK was however attributed to the design of the Census geography rather than any natural propensity for the population to have divergent characteristics.

The homogeneity of the cluster assignment across the three hierarchical levels of the 2011 OAC and the variation between clusters allowed instances of atypicality to be identified. This indicated that two 2011 OAC Supergroups: 'Constrained City Dwellers' and 'Cosmopolitans' contain the majority of atypical OAs and SAs in the UK. The results from this analysis indicate that there is variation in the extent to which all areal units assigned to a cluster will conform to its average characteristics, with some clusters having an increased propensity to contain atypical OAs and SAs. These variations are important and the use of SED values to identify these occurrences provides an indication of the complexity of the 2011 OAC.

It was important to assess the physical and social environmental change that has taken place in the UK since 2001, and how it was represented in the 2011 OAC. It was demonstrated that changes in the built-up environment in Basingstoke and the social make-up in Southampton since 2001 are reflected in the 2011 OAC Supergroups and Groups. This means users can have greater confidence in the 2011 OAC proving a useful representation of the UK. The two case studies were however limited to small geographical areas in the South East of England. As such, while it seems likely change will have been incorporated into the 2011 OAC, there cannot be any guarantee this is true for other parts of the UK.

The final validation exercise was ground-truthing the classification. This was performed by a number of students from UCL who visited different areas in London to assess if the

assigned 2011 OAC Supergroup provided the best description of the location. The results indicated that a greater number of students agreed that the Supergroup assignment to an area was the best option available – suggesting that assignment of a 2011 OAC cluster to an OA or SA will be correct more often than it is wrong. The limited geographical coverage and the assessment of only the Supergroup level of the 2011 OAC does however mean that the applicability of the results to all hierarchical levels of the classification and to areas outside of London is unknown.

The validation of the 2011 OAC indicates that it is a robust classification that consists of distinct clusters which incorporate change in the UK since 2001 and is correct more often than it is wrong. The scale of the classification does however mean that no validation exercise can encompass its entirety. These conclusions must therefore be tempered by the fact the dynamic nature of the UK means that inevitably the 2011 OAC will perform better in certain scenarios and geographic locations.

v) To explore alternatives to using ancillary data sources to update geodemographic classifications that highlight the temporal stability, or otherwise, of resident populations.

This aim focussed on addressing the temporal stability of geodemographic classifications after their release. This was detailed in Chapter 5, which is based on Gale and Longley (2013). Census based geodemographics do not capture the dynamics of change in small areas over time, increasing the possibility of representations becoming less effective at describing the characteristics of areas. The current inability in open geodemographics to replicate the use of private ancillary data sources used by commercial systems to update their classifications meant alternatives needed to be considered. The alternative solution proposed was the use of mid-year population estimates (MYEs) and dwelling stock counts, two of the limited number of Open Data sources on an annual basis at the smallest geographical area, to provide the basis for uncertainty indicators.

Uncertainty indicators are based on the premise that change in the UK varies both geographically and over time. Examining the variances of change across the UK provides a more transparent account of how different areas exhibit different change characteristics, both spatially and temporally. Applying these indicators to the 2001 OAC showed for the majority of locations the original geodemographic assignment of the classification remained valid. The limited change to the population and dwelling stock dynamics across large areas of the UK between 2001 and 2011 provides empirical

evidence that temporal uncertainty indicators can be used to gauge the stability of geodemographic classifications over time. It can therefore be concluded that costly and time-consuming updating through ancillary sources used by commercial providers is only needed for some parts of the UK.

The successful creation of temporal uncertainty indicators for the 2001 OAC suggests they should be actively used by the 2011 OAC user base. The addition of such indicators to a classification like the 2011 OAC means users become aware of the need to investigate potentially uncertain areas using alternative data sources in order to better understand any change in an area's dynamics and make more informed decisions. This initial foray into researching the temporal stability of classifications like the 2001 OAC indicates, in the absence of more traditional, and perhaps unsystematic, updating techniques used by commercial operators, they will form an essential part of future open geodemographic classifications.

9.3. Applications of the 2011 OAC

In an era of austerity and cost restraint the 2011 OAC is a free alternative to the commercial systems available. Although not directly comparable to these systems, the 2011 OAC does have its advantages, such as a methodology open to analysis and critique that can be adapted for use in bespoke applications. How the classification is deployed will depend on the specific needs of the user, however, examining the previous applications of the 2001 OAC allows generalisations to be made about its use.

The previous OAC User Group website (www.areaclassification.org.uk) details a number of different case studies of organisations who used the 2001 OAC. Local authorities for example have been keen users of the 2001 OAC, with Cambridgeshire County Council coding their Place Survey with the 2001 OAC, as discussed in Section 2.7.

The primary academic interest in the 2001 OAC has focused on its cross tabulation – tabulating it with other data sources to aid geodemographic analysis (Brunsdon et al., 2011; Singleton et al., 2012). It can therefore be expected that similar actions will be performed with the 2011 OAC. The linking of small area attributes from survey data, such as Understanding Society or the Crime Survey for England and Wales, to the 2011 OAC has the potential to add value and offer greater insight into the findings presented in these datasets.

An example of coding the 2011 OAC to survey data is shown in Figure 9.1. This shows one answer to the 'Taking everything into account, how good a job do you think the police in London as a whole are doing?' question from the Metropolitan Police Service's (MPS) Public Attitude Survey (PAS). It gives an indication of how the attitudes of the residents who live in the eight different 2011 OAC Supergroups differ.

Despite the 2001 OAC's open methodology, there has been only one documented adaption of it in the literature – the 2001 London Output Area Classification (2001 LOAC), which was discussed in Section 2.7. The 2001 LOAC created by Petersen et al. (2011) is a bespoke classification that modified the 2001 OAC methodology for a London only setting. This however was only conceived at the Supergroup level, limiting its scope when compared to the 2001 OAC. Other bespoke classifications, such as that created for Kingston upon Hull (Feldman, 2011), have utilised their own custom methodologies. The creation of a bespoke classification for Kingston upon Hull suggests that there is a demand for such outputs, with the open and transparent methodology of the 2011 OAC making such adaptions more straightforward.

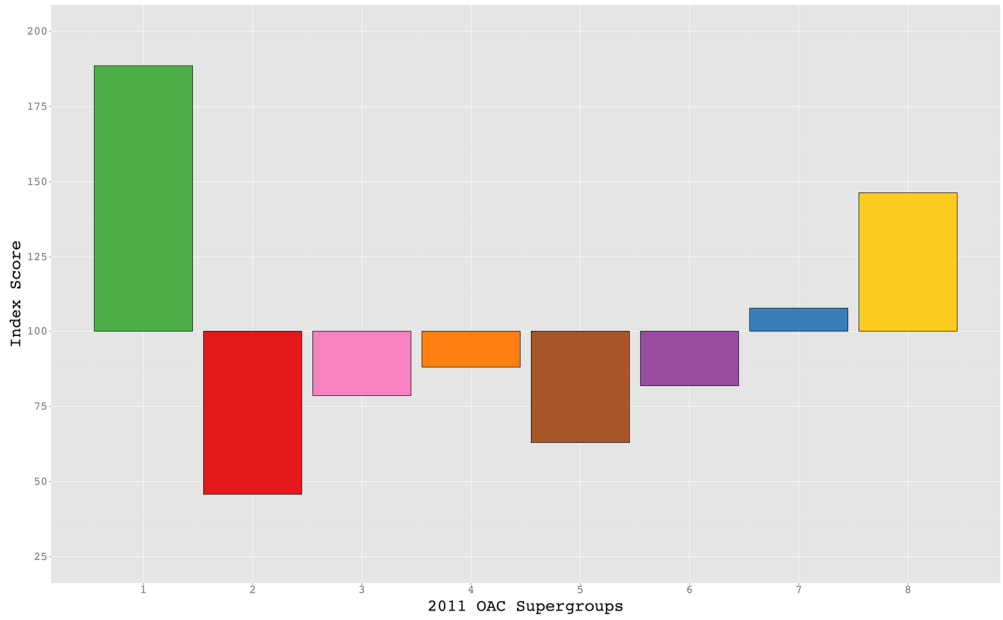


Figure 9.1: 2011 OAC Supergroup index scores for respondents who answered 'very poor' to the 'taking everything into account, how good a job do you think the police in London as a whole are doing?' question from the Metropolitan Police Service's Public Attitude Survey

The potential for the 2011 OAC's methodology to be adapted to create bespoke classifications has led to the Greater London Authority (GLA) commissioning the creation of a 2011 London Output Area Classification (2011 LOAC). Despite the 2011 OAC offering a better representation of London compared to the 2001 OAC, it was still deemed necessary by the GLA to commission the 2011 LOAC to provide the level of detail required for use in London only applications. Figure 9.2 shows the distribution of the named Supergroups in London. A key feature of the 2011 LOAC is the adoption of a hierarchical structure, similar to that of the 2011 OAC. In the case of the 2011 LOAC this hierarchical structure has been limited to 8 Supergroups and 19 Groups with a Subgroup level deemed unnecessary.

The benefits of the 2011 OAC therefore go beyond simply offering an alternative to the commercial systems that is free. The flexibility that the open and transparent methodology provides for users, offers them the opportunity to create their own classifications. As the majority of the operations are performed using open source programs and code it makes such adaptations easier than were previously possible. The 2011 LOAC is an example of how this approach can simplify the creation of new classifications. The benefits however are not limited to simply recreating the 2011 OAC at different regional scales, or only being of use to local government. It is conceivable that commercial companies that hold their own detailed records could incorporate these with other freely available data sources to create their own classification. This therefore would mean they could be self-reliant for their geodemographic requirements.

There can be no guarantee how widely the 2011 OAC will be adopted or what applications it might be used for. It can however be stated that the decisions made in the process of building the classification place it in a unique position within the wider geodemographic market in the UK. The fact that the classification is free is likely to be an appealing attribute for users, combined with the fact it can be used 'as is'. The open methodology, and in particular reliance on open source software separates the 2011 OAC from both the commercial systems and the 2001 OAC. The ability for users to adapt the classification to custom geographical regions or with alternative data sources is the unique selling point of the 2011 OAC and any applications that either use it directly or are adapted from it.

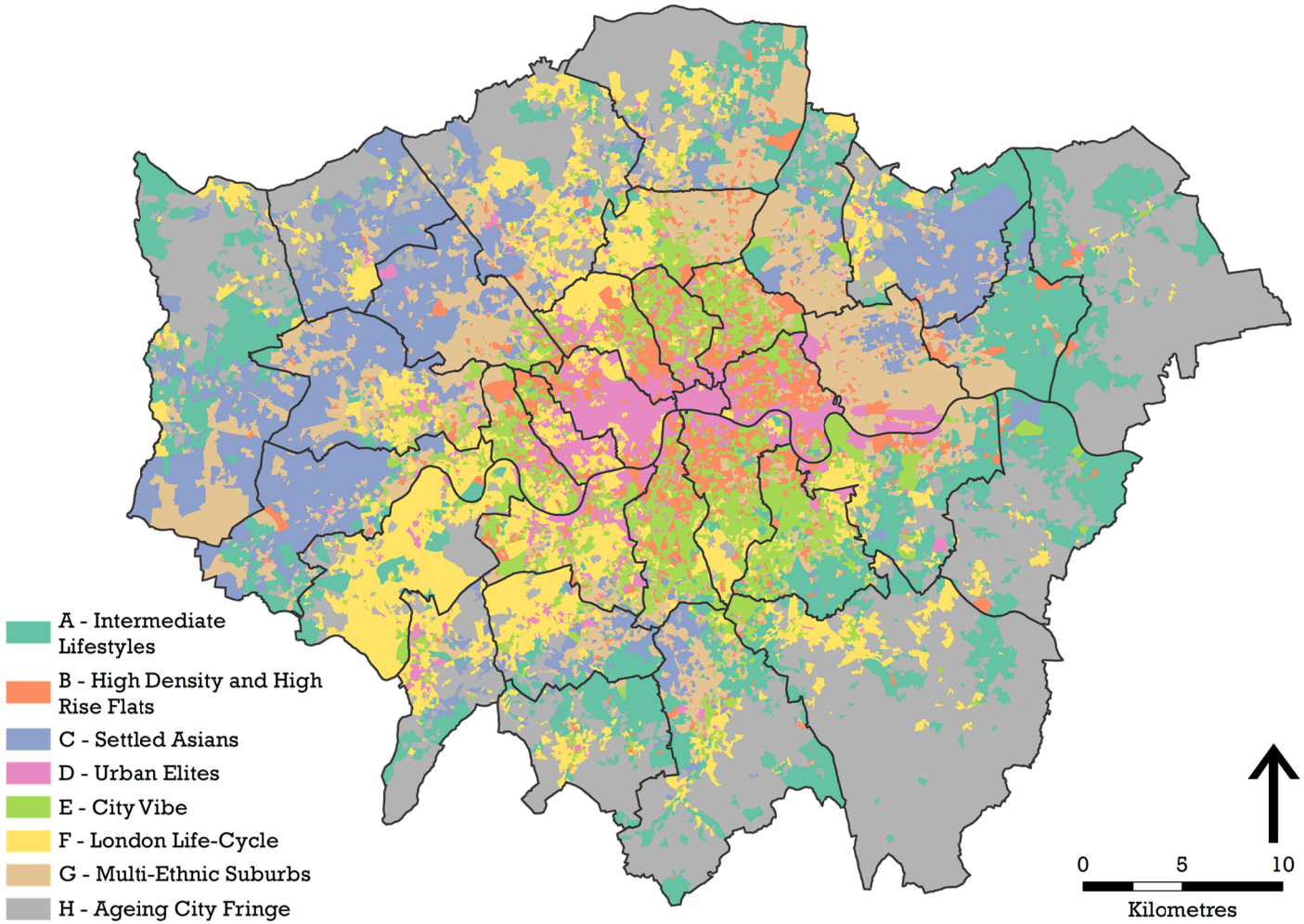


Figure 9.2: A choropleth map of the 2011 LOAC Supergroups

9.4. The lifespan of the 2011 OAC

The lifespan of the 2011 OAC is difficult to judge. The uncertainties that exist with the future of the UK Census, as discussed in Section 3.4, mean there is no guarantee the same extent of small-scale statistics in the UK will be available in the future. It is possible that the 2011 OAC will be the last open geodemographic classification of the UK constructed from the smallest area data, suggesting an extended lifespan as without a decennial census it will not be directly replaced. Results however from the 2011 OAC user engagement, discussed in Chapter 4, suggest that over time an increasing number of 2001 OAC users stopped using the classification. If this pattern were to repeat with the 2011 OAC it would therefore be logical to expect the classification to be most widely used at the point of release, and then over the subsequent years have a reduction in users.

If no replacement classification were possible in the 2020s it would be difficult to persuade any remaining users to continue using the 2011 OAC. Responses from the 2011 user engagement suggest users would consider an unmodified 2011 OAC to be an older, inferior product when compared to the commercial offerings by that point in its lifecycle. Chapter 5 proposes a methodology that can be used to understand the temporal uncertainty of geodemographic classifications. Taking advantage of Open Data sources discussed in Section 3.5 provides an indication of how change in the UK's population characteristics and physical characteristics are not constant over time, and the impact this has on geodemographic classification assignments. In the age of Open Data, temporal uncertainty measures can be considered an important part of any open geodemographic classification. The limitations of the technique and the inability to refresh classifications, similar to the commercial systems, mean a reliance on Census data remains.

The future of the 2011 OAC seems limited to creation of temporal uncertainty measures to provide an indication of the relative stability of its geodemographic assignments. The direction future open geodemographic classifications take seems likely to be influenced by what happens to the UK Census. This reliance on one particular data source makes open geodemographics susceptible to any change in government policy as the Beyond 2011 program has shown. Future open geodemographics research therefore has to focus on shifting attention to using the increasing number of Open Data sources.

The Economic and Social Research Council (ESRC) funded the 'Using Secondary Data to Measure, Monitor and Visualise Spatio-Temporal Uncertainties in Geodemographics' project at the University of Liverpool is the first step towards the integration of Open

Data sources into small level open geodemographic classifications. The project aims to improve on the temporal uncertainty techniques discussed in Chapter 5 by implementing a methodology for screening small areas of residential structures over time. Additionally, creating temporal measures of uncertainty and, crucially, providing updates to existing geodemographic classifications.

Research into the future development of small area open geodemographics without complete reliance on the UK Census needs to take into consideration the current limitations that exist, primary the availability of Open Data sources. As discussed in Section 3.5, there are currently very few Open Data sources that are available at Output Area (OA) or Small Area (SA) level. This means without modelling data, there is no way to introduce these sources into a geodemographic classification created at OA and SA level (such as the 2011 OAC). As such, there is a need to interweave frequently updated Open Data sources that are less likely to be available at OA and SA level with existing geodemographic classifications built at this level of granularity.

Future data availability and the level of granularity at which they are released is difficult to predict. Based on the current data landscape in the UK it appears the future of open geodemographic classifications, at least those that can be updated on a regular basis, is at coarser geographies. This mirrors the options available for the UK Census to either provide data less often but at a finer level of granularity, or more often at a coarser level of geography.

The combination of increasing numbers of Open Data sources, even if they are at coarser levels of geography, and frequently updated commercial systems is likely to impact what users expect from an open geodemographic classification. No longer will a decade long gap between releases be seen as acceptable. The likely outcome is that future research in open geodemographics will continue to use the most recent Census. The unparalleled breadth and quality of the data, discussed in Section 3.2, means that in certain cases no equivalent source exists in Open Data. The continued research into temporal uncertainty measures provides a framework for assessing the extent to which a geodemographic classification would need to be updated. Research should therefore be focused on identifying the Open Data sources that do exist that can supplement the Census data and provide temporal updates to geodemographic classifications.

The future of the 2011 OAC is dependent on a number of factors. Although regularly updating the classification in a similar way to commercial systems would be the ultimate ambition, such comprehensive procedures are not currently possible due to the limited range and geographical level of Open Data available. The extent to which this position will change over time will depend on future research into the use of Open Data with open geodemographics and the data made available. The alternative of using temporal uncertainty measures will provide an indication of the continuing relevance of the 2011 OAC in the future. The extent to which such techniques encourage continuing use of the 2011 OAC will be an interesting research topic in future years.

9.5. Concluding Comments

The 2011 OAC is an important tool in understanding the complexities of the UK's resident population. The division of the population into one of 8 Supergroups, 26 Groups and 76 Subgroups provides a clear and easy way of interpreting the socio-demographics of the UK. As the successor to the 2001 OAC, it takes advantage of computational advances over the past 10 years to provide not only an updated representation of the country, but also a blueprint that future geodemographic classifications can follow. This is made possible by the use of openly available data alongside open source software which provides a number of interesting research opportunities to continue advancing the open geodemographics agenda.

The pre-existing user awareness of the 2001 OAC and the current economic climate should encourage the wider use of the 2011 OAC compared to its predecessor. The open and transparent methodology gives the 2011 OAC an advantage over any commercial systems, appealing to users wishing to understand the underlying processes of the classification. This means the 2011 OAC is unique within the wider UK geodemographics market, although this alone does not guarantee its success. Although the creation of the 2011 OAC is an important milestone, it will be the continued use of the classification over a number of years that will be the true accomplishment. Future research in open geodemographics will be key to the longevity of the 2011 OAC, either through the creation of temporal uncertainty measures or regular updates.

This project has delivered a geodemographic classification that has made methodological advances beyond those of its predecessor. The use of open source software and freely accessible code provides users, with differing degrees of experience

with geodemographics, an opportunity to explore and utilise the classification. It is because of this the 2011 OAC can be considered a step forward for open geodemographics.

References

Abbott, O. (2009) '2011 UK Census Coverage Assessment and Adjustment Methodology', *Population Trends*, 137, pp. 25–32.

Adnan, M., Lansley, G. and Longley, P. A. (2013) 'A geodemographic analysis of the ethnicity and identity of Twitter users in Greater London', In *Proceedings of the 21st GIS Research UK Annual Conference*, Liverpool, UK.

Adnan, M., Longley, P. A., Singleton, A. D. and Brunson, C. (2010) 'Towards Real-Time Geodemographics: Clustering Algorithm Performance for Large Multidimensional Spatial Databases', *Transactions in GIS*, 14(3), pp. 283–297.

Aitchison, J. and Brown, J. A. C. (1957) *The lognormal distribution*, Cambridge, Cambridge University Press.

Akodjènou-Jeannin, M.-I., Salamatian, K. and Gallinari, P. (2007) 'Flexible Grid-Based Clustering', In Kok, J. N., Koronacki, J., Mantaras, R. L. de, Matwin, S., Mladenič, D., and Skowron, A. (eds.), *Knowledge Discovery in Databases: PKDD 2007*, Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp. 350–357.

Aldenderfer, M. S. and Blashfield, R. K. (1984) *Cluster analysis*, Thousand Oaks, California, Sage.

Aldstadt, J. and Getis, A. (2006) 'Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters', *Geographical Analysis*, 38(4), pp. 327–343.

Anderberg, M. R. (1973) *Cluster Analysis for Applications*, New York, Academic Press.

Arabie, P., Hubert, L. J. and De Soete, G. (eds.) (1996) *Clustering and Classification*, World Scientific.

Arnold, B. C. (2008) 'The Lorenz curve: Evergreen after 100 years', In Betti, G. and Lemmi, A. (eds.), *Advances on Income Inequality and Concentration Measures*, Routledge.

Ashby, D. I. and Longley, P. A. (2005) 'Geocomputation, Geodemographics and Resource Allocation for Local Policing', *Transactions in GIS*, 9(1), pp. 53–72.

Azuma, D. and Monleon, V. J. (2011) 'Differences in forest area classification based on tree tally from variable- and fixed-radius plots', *Canadian Journal of Forest Research*, 41(1), pp. 211–214.

Bailey, S., Charlton, J., Dollamore, G. and Fitzpatrick, J. (1999) 'Which authorities are alike?', *Population Trends*, 98, pp. 29–41.

Bailey, S., Charlton, J., Dollamore, G. and Fitzpatrick, J. (2000) 'Families, groups and clusters of local and health authorities of Great Britain: Revised for authorities in 1999', *Population Trends*, 99, pp. 37–52.

Bailey, T. C. and Gatrell, A. C. (1995) *Interactive spatial data analysis*, Longman Scientific & Technical.

Batey, P. W. J. and Brown, P. J. B. (1995) 'From Human Ecology to Customer Targeting: the Evolution of Geodemographics', In *GIS for Business and Service Planning*, Cambridge, GeoInformation International, pp. 77–103.

Batty, M. and Longley, P. A. (1986) 'The fractal simulation of urban structure', *Environment and Planning A*, 18(9), pp. 1143 – 1179.

Benton, P., Teague, A., Calder, A. and Naylor, J. (2013) 'Beyond 2011 - A new paradigm for population statistics?', In *59th ISI World Statistics Congress*, Hong Kong, China, [online] Available from: http://www.statistics.gov.hk/wsc/IPS027-P3-S.pdf (Accessed 2 February 2014).

Berry, M. J. A. and Linoff, G. S. (1996) *Data Mining Techniques: For Marketing, Sales, and Customer Support*, Wiley.

Bezdek, J. C. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*, Norwell, USA, Kluwer Academic Publishers.

Birkin, M. (1995) 'Customer targeting, Geodemographics and lifestyle approaches', In Longley, P. A. and Clarke, G. (eds.), *GIS for Business and Service Planning*, Cambridge, GeoInformation International.

Blake, M. and Openshaw, S. (1994) *GB Profiles: A User Guide*, Leeds, University of Leeds.

Blake, M. and Openshaw, S. (1995) *Selecting variables for small area classifications of 1991 UK census data*, Leeds, School of Geography, University of Leeds.

Blanchflower, D. G. and Lawton, H. (2008) *The Impact of the Recent Expansion of the EU on the UK Labour Market*, IZA Discussion Paper, Institute for the Study of Labor (IZA), [online] Available from: http://ideas.repec.org/p/iza/izadps/dp3695.html (Accessed 2 March 2014).

Bolton Metropolitan Borough Council (2005) '2001 Census Topic Report: Households and Housing', Bolton Metropolitan Borough Council, [online] Available from: http://www.bolton.gov.uk/sites/DocumentCentre/Documents/Census%20Topic%20Report%20-%20Housing.pdf (Accessed 20 September 2013).

Box, G. E. P. and Cox, D. R. (1964) 'An Analysis of Transformations', *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2), pp. 211–252.

Brewer, C. A. (1994) 'Color Use Guidelines for Mapping and Visualization', In MacEachren, A. M. and Taylor, D. R. F. (eds.), *Visualization in Modern Cartography*, New York, USA, Elsevier Science, pp. 123–147.

Brunsdon, C., Longley, P., Singleton, A. and Ashby, D. (2011) 'Predicting participation in higher education: a comparative evaluation of the performance of geodemographic classifications', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(1), pp. 17–30.

Bryman, A. (2008) *Social Research Methods*, Oxford, Oxford University Press.

Burbidge, J. B., Magee, L. and Robb, A. L. (1988) 'Alternative Transformations to Handle Extreme Values of the Dependent Variable', *Journal of the American Statistical Association*, 83(401), pp. 123–127.

Butler, T. and Robson, G. (2003) *London Calling: The Middle Classes and the Remaking of Inner London*, Oxford, Berg.

CACI (2009) 'Proud to be different – London found to be nothing like rest of UK', [online] Available from: http://www.caci.co.uk/395.aspx (Accessed 15 July 2013).

CACI (2013a) 'Acorn technical guide', CACI Ltd., [online] Available from: http://acorn.caci.co.uk/downloads/Acorn-Technical-document.pdf (Accessed 15 July 2013).

CACI (2013b) 'Acorn user guide', CACI Ltd., [online] Available from: http://acorn.caci.co.uk/downloads/Acorn-User-guide.pdf (Accessed 15 July 2013).

CACI (2013c) 'Client Testimonials', [online] Available from: http://www.caci.co.uk/testimonials.aspx (Accessed 15 July 2013).

Cambridgeshire County Council (2012) 'Community insight', [online] Available from: http://www.cambridgeshire.gov.uk/business/research/Social+Classification.htm (Accessed 7 August 2013).

Carpenter, J. and Watts, P. (2013) *Assessing the value of OS OpenData to the economy of Great Britain - Synopsis*, Ordnance Survey, [online] Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/207692/bis-13-950-assessing-value-of-opendata-to-economy-of-great-britain.pdf (Accessed 23 August 2013).

Chambers, R. and Tzavidis, N. (2006) 'M-quantile Models for Small Area Estimation', *Biometrika*, 93(2), pp. 255–268.

Charlton, M., Openshaw, S. and Wymer, C. (1985) 'Some new classifications of census enumeration districts in Britain: a poor man's ACORN', *Journal of Economic and Social Measurement*, 13(1), pp. 69–96.

Cockings, S., Harfoot, A., Martin, D. and Hornby, D. (2011) 'Maintaining existing zoning systems using automated zone-design techniques: methods for creating the 2011 Census output geographies for England and Wales', *Environment and Planning A*, 43(10), pp. 2399–2418.

Collins, J., Elliot, D., Walker, S., Watson, J. and Marques dos Santos, M. (2010) '2007 Census Test: The effects of including questions on income and implications for the 2011 Census', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/the-2011-census-project/2007-test/income-evaluation/2007-test-income-question-evaluation-report.pdf (Accessed 20 August 2013).

Congdon, P. D. (2010) *Applied Bayesian Hierarchical Methods*, CRC Press.

Cox, A. W., Lees, F. P. and Ang, M. L. (1990) *Classification of Hazardous Locations*, Institution of Chemical Engineers.

Dag, O., Asar, O. and Ilk, O. (2014) 'A Methodology to Implement Box-Cox Transformation When No Covariate is Available', *Communications in Statistics - Simulation and Computation*, 43(7), pp. 1740–1759.

Deloitte (2012) 'Open Growth. Stimulating demand for open data in the UK', Deloitte, [online] Available from: http://www.deloitte.com/assets/Dcom-UnitedKingdom/Local%20Assets/Documents/Market%20insights/Deloitte%20Analytics/uk-da-open-growth.pdf (Accessed 23 August 2013).

Department for Communities and Local Government (2011a) *The English Indices of Deprivation 2010*, Neighbourhoods Statistical Release, [online] Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/6871/1871208.pdf (Accessed 2 August 2011).

Department for Communities and Local Government (2011b) *Dwelling Stock Estimates: 2011, England*, Housing Statistical Release, [online] Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/6868/2039750.pdf (Accessed 10 February 2013).

DeWalt, K. M. and DeWalt, B. R. (2002) *Participant Observation: A Guide for Fieldworkers*, Rowman Altamira.

Dignan, T., IJpelaar, J., Marshall, D. and Watson, C. (2010) 'Small Area Population Estimates for Northern Ireland (2008)', *NISRA Occasional Paper*, Number 30.

Dixon, C. J. and Leach, B. (1978) *Questionnaires and interviews in geographical research*, Concepts and techniques in modern geography, Norwich, England, Geo Abstracts.

Dugmore, K. (2009) 'Information collected by Commercial Companies: What might be of value to ONS?', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/method-quality/imps/updates-and-reports/historical-updates-and-releases/updates-and-releases-from-2010/information-collected-by-commercial-companies--what-might-be-of-value-to-ons-.pdf (Accessed 20 August 2013).

Dugmore, K., Furness, P., Leventhal, B. and Moy, C. (2011) 'Beyond the 2011 Census in the United Kingdom: with an international perspective', *International Journal of Market Research*, 53(5), pp. 619–650.

Estivill-Castro, V. and Lee, I. (2000) 'Amoeba: Hierarchical Clustering Based On Spatial Proximity Using Delaunaty Diagram', In *Proceedings of the 9th International Symposium on Spatial Data Handling*, Beijing, China.

Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011) *Cluster Analysis*, 5th ed. Wiley Series in Probability and Statistics, Chichester, UK, Wiley.

Experian (2010) 'Optimise the value of your customers and locations, now and in the future: Mosaic UK - the consumer classification of the United Kingdom', [online] Available from: http://www.experian.co.uk/assets/business-strategies/brochures/mosaic-uk-2009-brochure-jun10.pdf (Accessed 15 July 2013).

Experian (2013) 'Mosaic UK - unique consumer classification based on in-depth demographic data: Client testimonials', [online] Available from: http://www.experian.co.uk/business-strategies/mosaic-uk.html#tabs-5 (Accessed 15 July 2013).

Fay, R. E. and Herriot, R. A. (1979) 'Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data', *Journal of the American Statistical Association*, 74(366), pp. 269–277.

Feldman, S. (2011) 'Developing a "free" customer classification system – The Alternative Approach!', [online] Available from: http://www.agi.org.uk/lps/2011/5/6/developing-a-free-customer-classification-system-the-alterna.html (Accessed 24 March 2014).

Feng, Z. and Flowerdew, R. (1998) 'Fuzzy geodemographics: a contribution from fuzzy clustering methods', In Carver, S. (ed.), *Innovations in GIS 5*, London, Taylor and Francis.

Fisher, P. F. and Tate, N. J. (In Press) 'Modelling Class Uncertainty in the Geodemographic Output Area Classification', *Environment and Planning B: Planning and Design*.

Flowerdew, R. and Leventhal, B. (1998) 'Under the microscope', *New Perspectives*, 18, pp. 16–38.

Fotheringham, A. S. and Wong, D. W. S. (1991) 'The Modifiable Areal Unit Problem in Multivariate Statistical Analysis', *Environment and Planning A*, 23(7), pp. 1025–1044.

Gale, C. G. and Longley, P. A. (2012) 'Geodemographic Output Area Classifications for London, 2001-2011', In *Proceedings of the 20th GIS Research UK Annual Conference*, Lancaster, UK.

Gale, C. G. and Longley, P. A. (2013) 'Temporal Uncertainty in a Small Area Open Geodemographic Classification', *Transactions in GIS*, 17(4), pp. 563–588.

Gardner, S. D. (2005) 'Evaluation of the ColorBrewer color schemes for accommodation of map readers with impaired color vision', The Pennsylvania State University, College of Earth and Mineral Sciences.

Gastner, M. T. and Newman, M. E. J. (2004) 'Diffusion-based method for producing density-equalizing maps', *Proceedings of the National Academy of Sciences of the United States of America*, 101(20), pp. 7499–7504.

Gehlke, C. and Biehl, K. (1934) 'Certain effects of grouping upon the size of the correlation coefficient in census tract material', *Journal of the American Statistical Association*, 29(185), pp. 169–170.

Gini, C. (1912) *Variabilità e Mutuabilità*, Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche. C. Cuppini, Bologna.

Goder, A. and Filkov, V. (2008) 'Consensus Clustering Algorithms: Comparison and Refinement', In *Proceedings of the Tenth Workshop on Algorithm Engineering and Experiments*, San Francisco, California, USA.

Goodchild, M. F. (2007) 'Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0', *International Journal of Spatial Data Infrastructures Research*, 2(2), pp. 24–32.

Gordon, A. D. (1999) *Classification*, 2nd ed. Monographs on Statistics and Applied Probability, Chapman & Hall/CRC Press.

Goss, J. (1995) 'Marketing the new marketing: The strategic discourse of geodemographic information systems', In Pickles, J. (ed.), *Ground Truth*, New York, Guildford Press, pp. 130–170.

Gould, P. and White, R. (1974) *Mental Maps*, Harmondsworth, Penguin.

GROS (2010) 'Mid-Year Population Estimates for Scotland: Methodology Guide', General Register Office for Scotland, [online] Available from: http://www.gro-scotland.gov.uk/files2/stats/population-estimates/mid-year-pop-est-methodology.pdf (Accessed 2 October 2013).

GROS (2013) '2011 Census Output Areas, Local Characteristic and Detailed Characteristic Sectors', General Register Office for Scotland, [online] Available from: http://www.gro-scotland.gov.uk/files2/geography/2011-census/2011-census-geography-background-info.pdf (Accessed 2 October 2013).

Gunesh, R. (2005) 'Correlation Analysis', [online] Available from: http://pages.intnet.mu/cueboy/education/notes/statistics/pearsoncorrel.pdf (Accessed 8 August 2012).

Haklay, M., Singleton, A. and Parker, C. (2008) 'Web Mapping 2.0: The Neogeography of the GeoWeb', *Geography Compass*, 2(6), pp. 2011–2039.

Hall, P., Marshall, S. and Lowe, M. (2001) 'The Changing Urban Hierarchy in England and Wales, 1913-1998', *Regional Studies*, 35(9), pp. 775–807.

Haring, L. L. and Lounsbury, J. F. (1975) *Introduction to Scientific Geographical Research*, Iowa, USA, W. C. Brown.

Harper, G. and Mayhew, L. (2012) 'Applications of Population Counts Based on Administrative Data at Local Level', *Applied Spatial Analysis and Policy*, 5(3), pp. 183–209.

Harris, R., Sleight, P. and Webber, R. (2005) *Geodemographics, GIS and Neighbourhood Targeting*, London, Wiley.

Harrower, M. and Brewer, C. A. (2003) 'ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps', *The Cartographic Journal*, 40(1), pp. 27–37.

Hasan, M. A., Chaoji, V., Salem, S. and Zaki, M. J. (2009) 'Robust partitional clustering by outlier and density insensitive seeding', *Pattern Recognition Letters*, 30(11), pp. 994–1002.

HM Government (2011) *Making Open Data Real: A Public Consultation*, [online] Available from: http://data.gov.uk/sites/default/files/Open%20Data%20consultation%20August%202011.pdf (Accessed 22 August 2013).

Holt, D. B. (1998) 'Does Cultural Capital Structure American Consumption?', *Journal of Consumer Research*, 25(1), pp. 1–25.

House of Commons Treasury Committee (2008) *Counting the population*, Eleventh Report of Session 2007–08, HMSO, [online] Available from: http://www.publications.parliament.uk/pa/cm200708/cmselect/cmtreasy/183/183.pdf (Accessed 22 August 2013).

Hoyt, H. (1939) *The structure and growth of residential neighborhoods in American cities*, Washington D.C., USA, Federal Housing Administration, [online] Available from: http://archive.org/details/structuregrowtho00unitrich (Accessed 17 October 2013).

IJpelaar, J., Marshall, D., Paul, S. and Moylan, K. (2011) 'Quality Report / User Guide – Northern Ireland Population Estimates', *NISRA Occasional Paper*, Number 32.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999) 'Data clustering: a review', *ACM Computing Survey*, 31(3), pp. 264–323.

Jankowska, M., Aldstadt, J., Getis, A., Weeks, J. and Fraley, G. (2008) 'An AMOEBA procedure for visualizing clusters', In *Proceedings of GIScience*.

Jelinski, D. and Wu, J. (1996) 'The modifiable areal unit problem and implications for landscape ecology', *Landscape Ecology*, 11(3), pp. 129–140.

Johnson, D. R. and Creech, J. C. (1983) 'Ordinal Measures in Multiple Indicator Models: A Simulation Study of Categorization Error', *American Sociological Review*, 48(3), pp. 398–407.

Johnson, N. L. (1949) 'Systems of Frequency Curves Generated by Methods of Translation', *Biometrika*, 36(1-2), pp. 149–176.

Johnson, R. A. and Wichern, D. W. (2007) *Applied Multivariate Statistical Analysis*, 6th ed. Pearson Prentice Hall.

Jones, S. (1999) *Doing Internet Research: Critical Issues and Methods for Examining the Net*, London, Sage.

Kakwani, N. (2010) 'The Lorenz curve', In Blaug, M. and Lloyd, P. J. (eds.), *Famous Figures and Diagrams in Economics*, Edward Elgar Publishing.

Kaski, S. (1997) 'Data Exploration Using Self-Organizing Maps', Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82.

Kaufman, L. and Rousseeuw, P. J. (2005) *Finding Groups in Data*, 2nd ed. Chichester, UK, Wiley.

Kawulich, B. B. (2005) 'Participant Observation as a Data Collection Method', *Forum: Qualitative Social Research*, 6(2), [online] Available from: http://www.qualitative-research.net/index.php/fqs/article/view/466 (Accessed 4 March 2014).

Kohonen, T. (1984) *Self-organisation and Associative Memory*, Berlin, Germany, Spring-Verlag.

Kohonen, T. (1998) 'The self-organising map', *Neurocomputing*, 21, pp. 1–6.

Kovács, F., Legány, C. and Babos, A. (2005) 'Cluster Validity Measurement Techniques', *Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence*, pp. 18–19.

Van Laerhoven, K. (2001) 'Combining the Self-Organizing Map and K-Means Clustering for On-Line Classification of Sensor Data', In.

Large, A. and Brown, J. (2010) 'Estimating and Correcting for Over-count in the 2011 Census', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/statistical-methodology/estimating-and-correcting-for-over-count-in-the-2011-census.pdf (Accessed 21 August 2013).

Leydesdorff, L. and Bensman, S. (2006) 'Classification and powerlaws: The logarithmic transformation', *Journal of the American Society for Information Science and Technology*, 57(11), pp. 1470–1486.

Likert, R. (1932) 'A technique for the measurement of attitudes', *Archives of Psychology*, 22(140), p. 55.

Longley, P. A. (2005) 'Urban Studies', In Kempf-Leonard, K. (ed.), *Encyclopaedia of Social Measurement*, San Diego, Elsevier, pp. 921–926.

Longley, P. A., Cheshire, J. A. and Mateos, P. (2011) 'Creating a regional geography of Britain through the spatial analysis of surnames', *Geoforum*, 42, pp. 506–516.

Longley, P. A., Goodchild, M., Maguire, D. J. and Rhind, D. W. (2010) *Geographic Information Systems and Science*, 3rd ed. Wiley.

Longley, P. A. and Singleton, A. D. (2009) 'Classification Through Consultation: Public Views Of The Geography Of The E-Society.', *International Journal of Geographical Information Science*, 23(6), pp. 737–763.

Lorenz, M. O. (1905) 'Methods of Measuring the Concentration of Wealth', *Publications of the American Statistical Association*, 9(70), pp. 209–219.

Lorr, M. (1983) *Cluster Analysis for the Social Sciences*, San Francisco, California, USA, Jossey-Bass.

MacQueen, J. B. (1967) 'Some Methods for Classification and Analysis of Multivariate Observations', In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, pp. 281–297.

Mandelbrot, B. B. (1982a) 'Technical correspondence: comment on computer rendering of fractal stochastic models', *Communications of the ACM*, 25(8), pp. 581–583.

Mandelbrot, B. B. (1982b) *The Fractal Geometry of Nature*, New York, W.H. Freeman.

Marr, T. R. (1904) *Housing Conditions in Manchester & Salford: A Report Prepared for the Citizens' Association for the Improvement of the Unwholesome Dwellings and Surroundings of the People, with the Aid of the Executive Committee*, Sherratt and Hughes, University Press.

Martin, D. (1998) 'Optimizing census geography: the separation of collection and output geographies', *International Journal of Geographical Information Science*, 12(7), pp. 673–685.

Martin, D. (2000) 'Towards the Geographies of the 2001 UK Census of Population', *Transactions of the Institute of British Geographers*, 25(3), pp. 321–332.

Martin, D. (2002a) 'Geography for the 2001 Census in England and Wales', *Population Trends*, 108, pp. 7–15.

Martin, D. (2002b) 'Output Areas for 2001', In Rees, P. H., Martin, D., and Williamson, P. (eds.), *The Census Data System*, Chichester, UK, Wiley, pp. 37–47.

Martin, D. (2010) 'Understanding the social geography of census undercount', *Environment and Planning A*, 42(11), pp. 2753 – 2770.

Martin, D., Nolan, A. and Tranmer, M. (2001) 'The application of zone-design methodology in the 2001 UK Census', *Environment and Planning A*, 33(11), pp. 1949 – 1962.

Milligan, G. W. (1996) 'Clustering validation: Results and implications for applied analyses', In Arabie, P., Hubert, L. J., and De Soete, G. (eds.), *Clustering and Classification*, World Scientific.

Milligan, G. W. and Cooper, M. C. (1987) 'Methodological review: Clustering methods', *Applied Psychological Measurement*, 11, pp. 329–354.

Milligan, G. W. and Cooper, M. C. (1988) 'A study of standardisation of variables in cluster analysis', *Journal of Classification*, 5, pp. 181–204.

Monti, S., Tamayo, P., Mesirov, J. and Golub, T. (2003) 'Consensus Clustering: A Resampling Based Method for Class Discovery and Visualization of Gene Expression Microaray Data', *Machine Learning*, 52, pp. 91–118.

Morphet, C. S. (1993) 'The mapping of small-area census data - a consideration of the role of enumeration district boundaries', *Environment and Planning A*, 25(9), pp. 1267 – 1277.

Mouffron, M., Rousseau, F. and Zhu, H. (2008) 'Secure Two-Party Computation of Squared Euclidean Distances in the Presence of Malicious Adversaries', In Pei, D., Yung, M., Lin, D., and Wu, C. (eds.), *Information Security and Cryptology*, Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp. 138–152.

Ng, R. T. and Han, J. (1994) 'Efficient and Effective Clustering Methods for Spatial Data Mining', In *Proceedings of 20th International Conference on Very Large Data Bases*, Santiago de Chile, Chile, pp. 144–155.

Nguyen, N. and Caruana, R. (2007) 'Consensus Clusterings', In *Seventh IEEE International Conference on Data Mining*, Omaha, USA, pp. 607–612.

NISRA (2011) 'How are Population Estimates created? The Methodology', Northern Ireland Statistics and Research Agency, [online] Available from: http://www.nisra.gov.uk/archive/demography/population/midyear/mye_methodology.pdf (Accessed 2 October 2013).

NISRA (2012) 'Census 2011: Population and Household Estimates for Local Government Districts in Northern Ireland', Northern Ireland Statistics and Research Agency, [online] Available from: http://www.nisra.gov.uk/Census/pop_stats_bulletin_2_2011.pdf (Accessed 1 October 2013).

NISRA (2013) 'Small Areas for Northern Ireland', Northern Ireland Statistics and Research Agency, [online] Available from: http://www.nisra.gov.uk/archive/geography/SAPaperJan2013.doc (Accessed 14 August 2013).

NRS (2013a) '2011 Census: First Results on Population and Household Estimates for Scotland - Release 1C (Part Two)', National Records of Scotland, [online] Available from: http://www.scotlandscensus.gov.uk/documents/censusresults/release1c/rel1c2sb.pdf (Accessed 20 December 2013).

NRS (2013b) 'Beyond 2011: Newsletter Issue 1', National Records of Scotland, [online] Available from: http://www.gro-scotland.gov.uk/files2/beyond-2011/Newsletters/b2011-newsletter-april2013.pdf (Accessed 20 August 2013).

NRS (2013c) '2011 Census: First Results on Population and Household Estimates for Scotland - Release 1B', National Records of Scotland, [online] Available from: http://www.scotlandscensus.gov.uk/documents/censusresults/release1b/rel1bsb.pdf (Accessed 1 October 2013).

O'Brien, H. L. and Toms, E. G. (2008) 'What is user engagement? A conceptual framework for defining user engagement with technology', *Journal of the American Society for Information Science and Technology*, 59(6), pp. 938–955.

Ojo, A. A., Vickers, D. and Ballas, D. (2012) 'The Segmentation of Local Government Areas: Creating a New Geography of Nigeria', *Applied Spatial Analysis and Policy*, 5(1), pp. 25–49.

ONS (2004a) '2001 Census: Manchester and Westminster Matching Studies Full Report', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/population-and-migration/pop-ests/local-authority-population-studies/2001-census---manchester-and-westminster-matching-studies-full-report.pdf (Accessed 20 August 2013).

ONS (2004b) 'Analysis of data and evidence for Kensington and Chelsea', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/population-and-migration/pop-ests/local-authority-population-studies/local-authority-studies/kensington-and-chelsea.pdf (Accessed 20 August 2013).

ONS (2005) 'Census 2001: Quality Report for England and Wales', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/census/census-2001/design-and-conduct/review-and-evaluation/evaluation-reports/quality-report/census-2001-quality-report.pdf (Accessed 20 August 2013).

ONS (2010a) 'User Engagement Strategy', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/about-ons/consultations/closed-consultations/user-engagement-strategy/user-engagement-strategy---final-version.pdf (Accessed 2 September 2013).

ONS (2010b) 'Mid-year population estimates short methods guide', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/population-and-migration/pop-ests/mid-year-population-estimates-short-methods-guide.pdf (Accessed 2 October 2013).

ONS (2011) 'Methodology Note on production of Super Output Area Population Estimates', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/population-and-migration/pop-ests/methodology-note-on-production-of-super-output-area-population-estimates.pdf (Accessed 15 October 2012).

ONS (2012a) '2011 Census - Population and Household Estimates for England and Wales, March 2011', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/dcp171778_270487.pdf (Accessed 10 January 2013).

ONS (2012b) 'Ethnicity and National Identity in England and Wales 2011', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/dcp171776_290558.pdf (Accessed 20 August 2013).

ONS (2012c) *Census result shows increase in population of London as it tops 8 million*, Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/rel/mro/news-release/census-result-shows-increase-in-population-of-london-as-it-tops-8-million/pdflondonnr0712.pdf (Accessed 7 August 2013).

ONS (2012d) '2011 Census Coverage Survey Summary', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/census-coverage-survey-summary.pdf (Accessed 20 August 2013).

ONS (2012e) 'The 2011 Census Coverage Assessment and Adjustment Process', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/coverage-assessment-and-adjustment-process.pdf (Accessed 20 August 2013).

ONS (2012f) 'Response Rates in the 2011 Census', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/response-rates-in-the-2011-census.pdf (Accessed 15 July 2013).

ONS (2012g) 'Providing the online census', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/the-2011-census-project/2011-census-updates-and-evaluation-reports/2011-census-update--providing-the-online-census.pdf (Accessed 20 August 2013).

ONS (2012h) 'Overcount Estimation and Adjustment', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/overcount-estimation-and-adjustment.pdf (Accessed 21 July 2013).

ONS (2012i) 'Changes to Output Areas and Super Output Areas in England and Wales, 2001 to 2011', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/geography/products/census/report--changes-to-output-areas-and-super-output-areas-in-england-and-wales--2001-to-2011.pdf (Accessed 14 August 2013).

ONS (2012j) 'User Engagement on a new United Kingdom Output Area Classification - Consultation Document', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/new-uk-output-area-classification/user-engagement-on-a-new-united-kingdom-output-area-classification---consultation-document.pdf (Accessed 20 August 2013).

ONS (2012k) 'User Engagement on a new United Kingdom Output Area Classification - Summary of Responses', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/new-uk-output-area-classification/user-engagement-on-uk-output-area-classification.pdf (Accessed 20 August 2013).

ONS (2012l) '2011 Census: Population Estimates for the United Kingdom, 27 March 2011', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/dcp171778_292378.pdf (Accessed 20 September 2013).

ONS (2012m) 'Explaining the Difference between the 2011 Census Estimates and the Rolled-Forward Population Estimates', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/differences-between-2011-census-est-and-rolled-forward-pop-est.pdf (Accessed 2 October 2013).

ONS (2012n) 'Population Ageing in the United Kingdom, its Constituent Countries and the European Union', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/dcp171776_258607.pdf (Accessed 15 January 2014).

ONS (2013a) *Beyond 2011: Options Report*, Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/what-are-the-options-/beyond-2011-options-report--o1-.pdf (Accessed 4 August 2013).

ONS (2013b) 'A Beginner's Guide to UK Geography', *A Beginner's Guide to UK Geography*, [online] Available from: http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/output-area--oas-/index.html (Accessed 21 August 2013).

ONS (2013c) 'Beyond 2011: Progress Report – July 2013', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/news/reports-and-publications/beyond-2011-progress-report-july-2013-p3.pdf (Accessed 20 August 2013).

ONS (2013d) 'Internal Migration Estimates', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/method-quality/quality/quality-information/social-statistics/summary-quality-report-for-internal-migration.pdf (Accessed 2 February 2014).

ONS (2013e) '2011 Census Variable and Classification Information: Part 1', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/information-by-variable/part-1--variable-and-classification-introduction.pdf (Accessed 11 November 2013).

Openshaw, S. (1977) 'A Geographical Solution to Scale and Aggregation Problems in Region-Building, Partitioning and Spatial Modelling', *Transactions of the Institute of British Geographers*, 2(4), pp. 459–472.

Openshaw, S. (1983) 'Multivariate analysis of census data', In Rhind, D. W. (ed.), *A census user's handbook*, London, Methuen & Co.

Openshaw, S. (1984a) 'Ecological Fallacies and the Analysis of Areal Census Data', *Environment and Planning A*, 16(1), pp. 17–31.

Openshaw, S. (1984b) *The Modifiable Areal Unit Problem*, Concepts and Techniques in Modern Geography, Norwich, Geo Books.

Openshaw, S. (1989) 'Making Geodemographics More Sophisticated', *Journal of the Market Research Society*, 31(4), pp. 111–131.

Openshaw, S. (1994) 'Two exploratory space-time attribute pattern analysers relevant to GIS', In Fotheringham, S. and Rogerson, P. (eds.), *Spatial Analysis and GIS*, London, Taylor and Francis, pp. 83–104.

Openshaw, S. (1995) 'Marketing Spatial Analysis: A Review of Prospects and Technologies', In Longley, P. A. and Clarke, G. P. (eds.), *GIS for Business and Service Planning*, Geoinformation, Cambridge, Wiley, pp. 227–246.

Openshaw, S., Cullingford, D. and Gillard, A. (1980) 'A Critique of the National Classifications of OPCS/PRAG', *Town Planning Review*, 51(4), p. 421.

Openshaw, S. and Rao, L. (1995) 'Algorithms for reengineering 1991 Census geography', *Environment and Planning A*, 27(3), pp. 425 – 446.

Openshaw, S. and Taylor, P. J. (1979) 'A million or so correlation coefficients: three experiments on the modifiable areal unit problem', In Wrigley, N. (ed.), *Statistical methods in the spatial sciences*, London, Pion, pp. 127–144.

Orford, S., Dorling, D., Mitchell, R., Shaw, M. and Davey Smith, G. (2002) 'Life and death of the people of London: a historical GIS of Charles Booth's inquiry', *Health & Place*, 8, pp. 25–35.

Osamor, V. C., Adebiyi, E. F., Oyelade, J. O. and Doumbia, S. (2012) 'Reducing the Time Requirement of k-Means Algorithm', *PLoS ONE*, 7(12).

Osbourne, J. W. (2002) 'Notes on the Use of Data Transformation', *Practical Assessment, Research & Evaluation*, 8(6).

Parfitt, J. (1997) 'Questionnaire design and sampling', In Flowerdew, R. and Martin, D. (eds.), *Methods in Human Geography: a guide for students doing research projects*, Essex, England, Pearson, pp. 76–109.

Parsons, T. and Knight, P. G. (2005) *How to do your dissertation in geography and related disciplines*, 2nd ed. London, Routledge.

Peach, C. (1996) 'Does Britain have ghettoes?', *Transactions of the Institute of British Geographers*, 21, pp. 216–235.

Peeples, M. A. (2011) 'R Script for K-Means Cluster Analysis', [online] Available from: http://www.mattpeeples.net/kmeans.html (Accessed 2 February 2014).

Pence, K. (2006) 'The Role of Wealth Transformations: An Application to Estimating the Effect of Tax Incentives on Saving', *Contributions to Economic Analysis & Policy*, 5(1), pp. 1430–1430.

Petersen, J., Gibin, M., Longley, P., Mateos, P., Atkinson, P. and Ashby, D. (2011) 'Geodemographics as a tool for targeting neighbourhoods in public health campaigns', *Journal of Geographical Systems*, 13(2), pp. 173–192.

Pharoah, R. and Rowe, B. (2010) 'Why Westminster will prove "hard-to-count" in the 2011 census', ESRO, [online] Available from: http://www.westminster.gov.uk/workspace/assets/publications/2011-Census-Why-Westminster-wil-1285775755.pdf (Accessed 21 August 2013).

Plewe, B. (2002) 'The Nature of Uncertainty in Historical Geographic Information', *Transactions in GIS*, 6(4), pp. 431–456.

Plewis, I., Simpson, L. and Williamson, P. (2011) '2011: Independent Review of Coverage Assessment, Adjustment and Quality Assurance', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/the-2011-census-project/independent-assessments/independent-review-of-coverage-assessment--adjustment-and-quality-assurance/independent-review-final-report.pdf (Accessed 20 August 2013).

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*, Vienna, Austria, The R Foundation for Statistical Computing, [online] Available from: http://www.R-project.org (Accessed 14 March 2014).

Rao, J. N. K. (2005) *Small Area Estimation*, New York, USA, Wiley.

Raper, J., Rhind, D. W. and Shepherd, J. (1992) *Postcodes: The New Geography*, London, Longman.

Rees, P. H., Martin, D. and Williamson, P. (2002) 'Census data resources in the United Kingdom', In Rees, P. H., Martin, D., and Williamson, P. (eds.), *The Census Data System*, Chichester, UK, Wiley.

Robinson, W. S. (1950) 'Ecological Correlations and the Behavior of Individuals', *American Sociological Review*, 15(3), pp. 351–357.

Romesburg, C. (2004) 'Cluster Analysis for Researchers', *Cluster Analysis For Researchers*, p. 334.

Rousseeuw, P. J. (1987) 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics*, 20, pp. 53–65.

Sassen, S. (2001) *The Global City: New York, London, Tokyo*, Princeton University Press.

Shakespeare, S. (2013) *Shakespeare review: an independent review of public sector information*, Department for Business, Innovation & Skills, [online] Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/198752/13-744-shakespeare-review-of-public-sector-information.pdf (Accessed 22 August 2013).

Shelton, N., Birkin, M. H. and Dorling, D. (2006) 'Where Not to Live: a Geo-Demographic Classification of Mortality for England and Wales, 1981-2000', *Health and Place*, 12(4), [online] Available from: internal-pdf://shelton2006-0328823895/shelton2006.pdf.

Sheridan, J. and Tennison, J. (2010) 'Linking UK government data', In *19th International World Wide Web Conference*, Raleigh, North Carolina USA, [online] Available from: http://events.linkeddata.org/ldow2010/papers/ldow2010_paper14.pdf (Accessed 22 August 2013).

Shevky, E. and Williams, M. (1949) *The Social Areas of Los Angeles: Analysis and Typology*, Berkeley, University of California Press.

Shevsky, E. and Bell, W. (1955) *Social Area Analysis: Theory, Illustrative Application, and Computational Procedures*, Stanford, Stanford University Press.

Simpson, L. (2003) 'Are the Census outputs fit for purpose?', In *Royal Statistical Society/Office for National Statistics Census Conference 11-12*.

Simpson, S. (2002) 'Dealing with the census undercount', In Rees, P. H., Martin, D., and Williamson, P. (eds.), *The Census Data System*, Chichester, UK, Wiley.

Simpson, T. I., Armstrong, J. D. and Jarman, A. P. (2010) 'Merged consensus clustering to assess and improve class discovery with microarray data', *BMC Bioinformatics*, 11(590).

Singleton, A. D. (2010) 'The geodemographics of educational progression and their implications for widening participation in higher education', *Environment and Planning A*, 42(11), pp. 2560 – 2580.

Singleton, A. D. and Longley, P. A. (2009a) 'Geodemographics, visualisation, and social networks in applied geography', *Applied Geography*, 29(3), pp. 289–298.

Singleton, A. D. and Longley, P. A. (2009b) 'Creating Open Source Geodemographics - Refining a National Classification of Census Output Areas for Applications in Higher Education', *Papers in Regional Science*, 88(3), pp. 643–666.

Singleton, A. D. and Spielman, S. (Forthcoming) 'An Open Geodemographic Classification of the United States', *Annals of the Association of American Geographers*.

Singleton, A. D. and Spielman, S. (2013) 'The Past, Present and Future of Geodemographic Research in the United States and United Kingdom', *Professional Geographer*.

Singleton, A. D., Wilson, A. G. and O'Brien, O. (2012) 'Geodemographics and spatial interaction: an integrated model for higher education', *Journal of Geographical Systems*, 14(2), pp. 223–241.

Sleight, P. (2004) *Targeting Customers: How to Use Geodemographic and Lifestyle Data in Your Business*, 3rd ed. Henley-on-Thames, World Advertising Research Center Limited.

Slingsby, A., Dykes, J. and Wood, J. (2011) 'Exploring Uncertainty in Geodemographics with Interactive Graphics', *IEEE Transactions on Visualization and Computer Graphics*, 17(12), pp. 2545–2554.

Szekely, G. J. and Rizzo, M. L. (2005) 'Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method', *Journal of Classification*, 22(2), pp. 151–183.

The Boundary Commission for England (2013) '2013 Review of Parliamentary constituencies', The Boundary Commission for England, [online] Available from: http://consultation.boundarycommissionforengland.independent.gov.uk/wp-content/uploads/2013/07/Closure-report-Final-as-published.pdf (Accessed 20 August 2013).

The National Archives (2013) 'Open Government Licence v2.0', *Open Government Licence v2.0*, [online] Available from: https://www.nationalarchives.gov.uk/news/855.htm (Accessed 22 August 2013).

Tibshirani, R., Walther, G. and Hastie, T. (2000) 'Estimating the number of clusters in a dataset via the Gap statistic', 63, pp. 411–423.

Tobler, W. R. (1970) 'A Computer Movie Simulating Urban Growth in the Detroit Region', *Economic Geography*, 46, pp. 234–240.

Tranmer, M. and Steel, D. (2001) 'Using local census data to investigate scale effects', In Tate, N. J. and Atkinson, P. M. (eds.), *Modelling scale in geographical information science*, Chichester, UK, Wiley, pp. 105–122.

Tzavidis, N., Marchetti, S. and Chambers, R. (2010) 'Robust Estimation of Small-Area Means and Quantiles', *Australian & New Zealand Journal of Statistics*, 52(2), pp. 167–186.

UK Data Service (2013) 'Census aggregate data guide', [online] Available from: http://census.ukdataservice.ac.uk/use-data/guides/aggregate-data.aspx (Accessed 1 October 2013).

Unwin, D. J. (1996) 'GIS, spatial analysis and spatial statistics', *Progress in Human Geography*, 20(4), pp. 540–551.

Vargas-Silva, C. (2013) *Migration Flows of A8 and other EU Migrants to and from the UK*, The Migration Observatory at the University of Oxford, [online] Available from: http://www.migrationobservatory.ox.ac.uk/briefings/migration-flows-a8-and-other-eu-migrants-and-uk (Accessed 2 March 2014).

De Vaus, D. A. (2002) *Social Surveys*, London, Sage.

Vickers, D. W. (2003) 'The difficulty of linking two differently aggregated spatial datasets: using a look-up table to link postal sectors and 1991 Census enumeration districts', *Working Paper 03/2*, School of Geography, University of Leeds, [online] Available from: http://www.geog.leeds.ac.uk/fileadmin/documents/research/csap/wpapers/03-2.pdf (Accessed 12 July 2013).

Vickers, D. W. (2006) 'Multi-level Integrated Classifications Based on the 2001 Census', University of Leeds, Department of Geography.

Vickers, D. W. and Rees, P. H. (2007) 'Creating the UK National Statistics 2001 output area classification', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2), pp. 379–403.

Vickers, D. W. and Rees, P. H. (2011) 'Ground-truthing Geodemographics', *Applied Spatial Analysis and Policy*, 4(1), pp. 3–21.

Vickers, D. W., Rees, P. H. and Birkin, M. (2005) *Creating the National Classification of Census Output Areas: Data, Methods and Results*, School of Geography, University of Leeds, [online] Available from: http://www.geog.leeds.ac.uk/fileadmin/documents/research/csap/wpapers/05-2.pdf (Accessed 12 July 2013).

VOA (2008) 'Understanding your council tax banding', VOA, [online] Available from: http://www.voa.gov.uk/corporate/_downloads/pdf/VO7858_understanding_ct.pdf (Accessed 2 October 2013).

Voas, D. and Williamson, P. (2001) 'The diversity of diversity: a critique of geodemographic classification', *Area*, 33(1), pp. 63–76.

Walford, N. (2013) 'An Introduction to Geodemographic Classification', [online] Available from: http://pldocs.docdat.com/download/docs-97092/97092.doc (Accessed 15 July 2013).

Wallace, M. and Denham, C. (1996) *The ONS Classification of Local and Health Authorities of Great Britain*, Studies on Medical and Population Subjects, London, HMSO.

Ward, J. H. (1963) 'Hierarchical Grouping to Optimize an Objective Function', *Journal of the American Statistical Association*, 58(301), pp. 236–244.

Webber, R. J. (1975) *Liverpool social area study, 1971 Data*, Planning Research Applications Group Technical Paper, London, Centre for Environmental Studies.

Webber, R. J. (1977) *An Introduction to the National Classification of Wards and Parishes*, Planning Research Applications Group Technical Paper, London, Centre for Environmental Studies.

Webber, R. J. (1978) 'Making the Most of the Census for Strategic Analysis', *The Town Planning Review*, 49(3), pp. 274–284.

Webber, R. J. (2007) 'The Metropolitan Habitus: Its Manifestations, Locations, and Consumption Profiles', *Environment and Planning A*, 39(1), pp. 182–207.

Webber, R. J. and Craig, J. A. (1976) 'Which local authorities are alike?', *Population Trends*, 5, pp. 13–19.

Webber, R. J. and Craig, J. A. (1978) *Socio-Economic Classification of Local Authority Areas*, Studies in Medical and Population Subjects, Office of Population Censuses and Surveys.

Weiss, M. J. (2000) *The Clustered World: How We Live, What We Buy, and What It All Means About Who We Are*, Little Brown & Company.

Welsh Assembly Government (2004) 'Council Tax Revaluation and Rebanding 2005', Welsh Assembly Government, [online] Available from: http://wales.gov.uk/dsjlg/publications/localgov/counciltaxravalue2005/guide?lang=en (Accessed 2 October 2013).

Wrigley, N. (1995) 'Revisiting the modifiable areal unit problem and the ecological fallacy', In Cliff, A. D., Gould, P. R., Hoare, A. G., and Thrift, N. J. (eds.), *Diffusing Geography*, Oxford, Blackwell.

Xu, R. and Wunsch, D. C. (2009) *Clustering*, IEEE series on computational intelligence, Oxford, Wiley.

Zadeh, L. A. (1965) 'Fuzzy sets', *Information and control*, 8(3), pp. 338–353.

Zumbo, B. D. and Zimmerman, D. W. (1993) 'Is the selection of statistical methods governed by level of measurement?', *Canadian Psychology/Psychologie canadienne*, 34(4), pp. 390–400.

Appendix A

A.1. User Engagement on a new United Kingdom Output Area Classification response form

Appendix A provides a sample form used by the 38 respondents to the 2011 OAC user engagement.

User Engagement on a new United Kingdom Output Area Classification

February 2012

1. Introduction

A joint funded project has been set up by the Office for National Statistics (ONS) and University College London (UCL) to create a new Output Area Classification using 2011 Census data. The original classification was created for the ONS by Dr Daniel Vickers at the University of Leeds using 2001 Census data. The undertaking of the 2011 Census provides an opportunity for the 2001 Output Area Classification to be updated using 2011 Census data, and for the methodology used to create the OAC to be reviewed.

It is intended that following the creation of a 2011 OAC, that area classifications for higher geographies such as Super Output Areas/Data Zones and local authorities will also be revised in a similar way.

1.1 Topic and scope

To help with the construction of this new open-source classification, and to better understand user requirements for it, ONS and UCL would welcome your thoughts, expectations and requirements for this new geodemographic classification. By answering any or all of the questions below you can help to shape the proposed Output Area Classification for 2011. Please use the section after question 20 to address any relevant points that you think may not have been addressed by the other questions.

1.2 Who we are seeking views from

We would particularly like to hear from regular users of any or all of the existing Area Classifications, and in particular from users across the UK of the current 2001 Output Area Classification, and potential users of a 2011 Output Area Classification.

1.3 Consultation timetable

This consultation will run for six weeks from 17 February 2012 to 30 March 2012.

1.4 After the consultation

Responses will be analysed by ONS and UCL, and ONS will publish a summary of the comments made approximately one month after the user engagement closes. The response template asks whether or not you agree to your responses being made public.

1.5 How to respond

Interested parties are invited to respond using the template in **Appendix 1** by the closing date via email:

2011OAC@ons.gov.uk

or by post to:

Andy Bates, Regional and Local Division, Office for National Statistics, Segensworth Road, Fareham
PO15 5RR

1.6 Confidentiality and data protection

Information provided in response to this consultation, including personal information, may be subject to publication or release to other parties or to disclosure in accordance with the access to information regimes (these are primarily the Freedom of Information Act 2000 (FOIA), the Data Protection Act 1998 (DPA) and the Environmental Information Regulations 2004).

If you would like the information, including personal data, that you submit to be treated as confidential, please be aware that, under the FOIA, there is a statutory Code of Practice with which public authorities must comply and which deals, among other things, with obligations of confidence. In view of this it would be helpful if you could explain to us why you regard the information you have provided as confidential. If we receive a request for disclosure of the information we will take full account of your explanation, but we cannot guarantee that confidentiality can be maintained in all circumstances. Before we disclose any information that is personal to you, we will inform you in advance of any disclosure. An automatic confidentiality disclaimer generated by your IT system will not, of itself, be regarded as binding on the Office for National Statistics.

Please ensure that your response is clearly marked if you wish your response and name to be kept confidential. Confidential responses will be included in any summary of numbers of comments received and views expressed.

This page is intentionally blank.

Appendix 1: User Engagement response template

Interested parties are invited to respond using this template by the closing date via email to:

2011OAC@ons.gov.uk

or by post to:

Andy Bates, Regional and Local Division, Office for National Statistics, Segensworth Road, Fareham
PO15 5RR

Your name *(optional)*

Organisation and Role (*optional*)

Do you wish your responses to be kept confidential?

☐ *Yes*
☐ *No*

The current 2001 Output Area Classification

1. Do you know what the current 2001 Output Area Classification (2001 OAC) is?
   ☐ *Yes*
   ☐ *No*

2. Do you (or your organisation) **currently** use the 2001 OAC?
   ☐ *Yes*
   ☐ *No*

If you answered "Yes" how long have you (or your organisation) been using the 2001 OAC for?

☐ *1 year or less*
☐ *2 to 5 years*
☐ *Over 5 years*

If you answered 'Yes' then please go to Question 4.

3. If you answered "No" to Question 2 have you (or your organisation) **previously** used the 2001 OAC?

☐ *Yes*
☐ *No*

If you answered "Yes" how long ago did you (or your organisation) stop using the 2001 OAC and why?

☐ *1 year or less*
☐ *2 to 5 years*
☐ *Over 5 years*

Why did you stop using the 2001 OAC?

If you answered "No" why have you never used the 2001 OAC?

4. What alternative commercial geodemographic classifications do you (or your organisation) use?

Select all that apply

☐ *ACORN by CACI*
☐ *Mosaic by Experian*
☐ *People & Places P^2 by Beacon Dodsworth*
☐ *PersonicX by Acxiom*
☐ *Other (please specify)*

☐ *None – I/we do not use any commercial geodemographic classifications*

Please briefly explain why you either do or do not use commercial geodemographic classification products:

5. Please indicate the geographical coverage(s) you favour when using a geodemographic classification?
Select all that apply

☐ *UK*
☐ *Country*
☐ *Regional*
☐ *County*
☐ *Local Authority*
☐ *City, Town or Village*
☐ *District or Area of a City, Town or Village*

6. Would you welcome a new version of the 2001 OAC?
☐ *Yes*
☐ *No*

7. Should a new 2011 Output Area Classification (2011 OAC) be a general purpose classification (like the 2001 OAC), or should it focus on producing specialised variants (such as health, education, crime etc.)?
*Please select **one** option only*

☐ *General purpose*
☐ *Specialised variants*

8. Flexibility in specifying the variables that are to make up the 2011 OAC would open up a range of options for area classification using Open Government Data. Is it important to you that the 2011 OAC be directly comparable – in terms of similar census data being used to construct it - with the 2001 OAC?
☐ *Yes*
☐ *No*

If you answered 'No' then what are the other priorities that are important to you in the construction of the 2011 OAC?
☐ *Updateable*
☐ *Better variables*
☐ *Other (please specify)*

9. The 2001 OAC divides the population of the UK into 7 Supergroups, 21 Groups and 52 Subgroups. How would you describe this framework when using the 2001 OAC for your particular purposes?

☐ *Extremely limited*
☐ *Limited*
☐ *Satisfactory*
☐ *Good*
☐ *Excellent*

New for the 2011 Output Area Classification

10. Thinking about how you use and interpret the 2001 OAC, how useful do you think each to the following options would be to you for the 2011 OAC?
Please tick a number from 1 to 5 or 'Don't know' to indicate your view. The equally spaced scale ranges from 1 = Not at all useful to 5 = Extremely useful.

Maps in PDF (or similar) format that are not interactive

☐ *1* ☐ *2* ☐ *3* ☐ *4* ☐ *5* ☐ *Don't know*

Online interactive maps with clickable details (such as the one found here for the 2001 OAC **http://www.maptube.org/map.aspx?mapid=960**)

☐ *1* ☐ *2* ☐ *3* ☐ *4* ☐ *5* ☐ *Don't know*

Mapping against different backdrops (such as Google Maps or OpenStreetMap)

☐ *1* ☐ *2* ☐ *3* ☐ *4* ☐ *5* ☐ *Don't know*

Correlation tables (showing to what extent the variables within the classification correlate with each other)

☐ *1* ☐ *2* ☐ *3* ☐ *4* ☐ *5* ☐ *Don't know*

Bar graphs of the group's attributes (**click here** for an example)

☐ *1* ☐ *2* ☐ *3* ☐ *4* ☐ *5* ☐ *Don't know*

Radial plots of the group's attributes (**click here** for an example)

☐ *1* ☐ *2* ☐ *3* ☐ *4* ☐ *5* ☐ *Don't know*

11. Thinking about your own understanding of the existing 2001 OAC, how useful do you think each to the following options would be to you for the 2011 OAC?
Please tick a number from 1 to 5 or 'Don't know' to indicate your view. The equally spaced scale ranges from 1 = Not at all useful to 5 = Extremely useful.

Group Name

☐ *1* ☐ *2* ☐ *3* ☐ *4* ☐ *5* ☐ *Don't know*

Graphical Representation (radial plots and bar graphs)

☐ *1* ☐ *2* ☐ *3* ☐ *4* ☐ *5* ☐ *Don't know*

Group definitions (a written summary of the key characteristics of each group)

☐ *1* ☐ *2* ☐ *3* ☐ *4* ☐ *5* ☐ *Don't know*

Key points of characteristics you would expect to find in each group

☐ *1* ☐ *2* ☐ *3* ☐ *4* ☐ *5* ☐ *Don't know*

Written 'pen portraits' of typical households found within each group

☐ *1* ☐ *2* ☐ *3* ☐ *4* ☐ *5* ☐ *Don't know*

Written 'pen portraits' of typical housing and built environments found in each group

☐ *1* ☐ *2* ☐ *3* ☐ *4* ☐ *5* ☐ *Don't know*

12. Do you agree with the view that it would be helpful to adjust the composition of each group for different parts of the UK (so, for example, there might be separate classifications made for London, or Scotland)?
☐ *Strongly disagree*
☐ *Disagree*
☐ *Neither agree nor disagree*
☐ *Agree*
☐ *Strongly agree*

13. Please identify what, if any, extra features would you like the 2011 OAC to have when compared with the 2001 OAC:

Dissemination of the 2011 OAC

14. Which methods of dissemination for the 2011 OAC would you be most likely to use?
Select all that apply

☐ *Online interactive mapping (such as the one found here for the 2001 OAC **www.maptube.org/map.aspx?mapid=960**)*

☐ *Enhanced online interactive mapping – with additional features that allow for you to search and identify the national 2011 OAC for a ward, county, local authority or region for example.*

☐ *Microsoft Excel/CSV file(s) containing the 2011 OAC classification for each of the UKs 2011 Census Output Areas)*

☐ *Software to append the 2011 OAC codes to a list of postcodes provided by the user (similar to the OAC Coder available at **www.publicprofiler.org**)*

☐ *Digital Boundary Data (eg a shapefile – a computer readable map that would provide the outline of Output Areas along with the 2011 OAC data – this would require the use of GIS software such as ArcGIS or MapInfo)*

15. Other data sources could be used to give greater context to the 2011 OAC. Rather than contributing to the classification itself, these could be used to help visualise the 2011 OAC in different ways. What (if any) data sources would you like to be able to use alongside the final 2011 OAC output?
Select all that apply

☐ *Index of Multiple Deprivation (the LSOA, Data Zone or SOA which the Output Area lies within)*
☐ *Temporal data – to attempt to distinguish different periods in a day*
☐ *Other (please specify)*

Construction of the 2011 OAC

16. There are multiple levels of spatial resolution that data can be produced at (see Appendix 2: Glossary of Terms for further information). In addition to Output Areas are there any other spatial resolutions you believe would benefit from having their own classification?
Select all that apply

☐ *Lower Layer Super Output Areas (LSOAs) in England & Wales / Data Zones in Scotland / Super Output Areas (SOAs) in Northern Ireland*
☐ *Middle Layer Super Output Areas (MSOAs)*
☐ *Wards*
☐ *Local Authorities*
☐ *Counties*
☐ *Regions*
☐ *Other (please specify)*

17. The 2001 OAC uses only 2001 Census data in its construction. It has been suggested that, in addition to using 2011 Census data, it might be possible for the 2011 OAC to be enhanced with supplementary non-census open data sources, and updated periodically over time. Would you find this beneficial?
☐ *Yes*
☐ *No*
☐ *Don't know*

Please briefly explain the reasons for your answer:

18. It is unlikely that many open data sources will offer UK wide coverage. What extent of coverage do you believe is a minimum requirement for an acceptable general purpose OAC classification?
☐ *UK only*
☐ *Countrywide coverage for England, Wales, Scotland or Northern Ireland*
☐ *Regional (ie classifications pertaining to parts of England, Wales, Scotland or Northern Ireland)*
☐ *Local Authority*
☐ *Other (please specify)*

19. If the 2011 OAC could be updated with new data, how frequently should this be done?
☐ *Once a year*
☐ *Every two years*
☐ *Every three years or longer*

20. Change in the social, economic and demographic structure of areas in the UK occurs at different rates. Instead of updating the 2011 OAC it might be possible to use non-census sources to flag areas where population changes have occurred, enabling the user to recognise parts of the UK where the classification had probably become unreliable. Would you find this helpful?
☐ *Yes*
☐ *No*
☐ *Don't know*

Any Other Comments

Appendix 2: Glossary of Terms

2001 OAC

The 2001 Output Area Classification is a geodemographic classification that distils key results from the 2001 Census for the whole of the UK at a fine spatial level of granularity to indicate the character of local areas. It was created in collaboration between the Office for National Statistics (ONS) and the University of Leeds

2011 Census

The 2011 Census was a count of all people and households in the UK. It provides population statistics from a national to neighbourhood level for government, local authorities, business and communities. It was carried out by the Office for National Statistics (ONS) in England and Wales, the National Records of Scotland (NRS) in Scotland and the Northern Ireland Statistics & Research Agency (NISRA) in Northern Ireland.

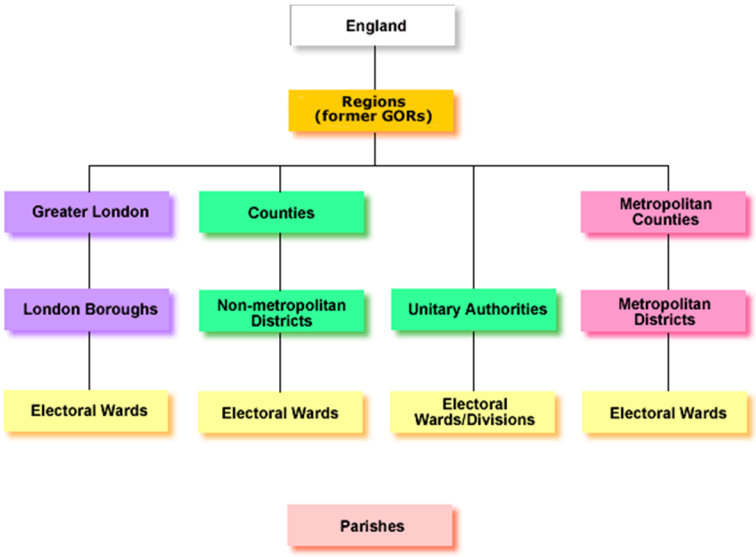
2011 OAC

The 2011 Output Area Classification is planned to be a geodemographic classification utilising data from the 2011 Census along with a variety of open-data sources to provide an indicator of local area characteristics. Particular focus will be on new modes of dissemination that better utilise Web technologies and new advances in GIS and geodemographics. It is being created as collaboration between the Office for National Statistics (ONS) and University College London (UCL).

Administrative Geographies

England

England does not have its own devolved parliament and is thus entirely subject to the administration of the UK Government in Westminster.



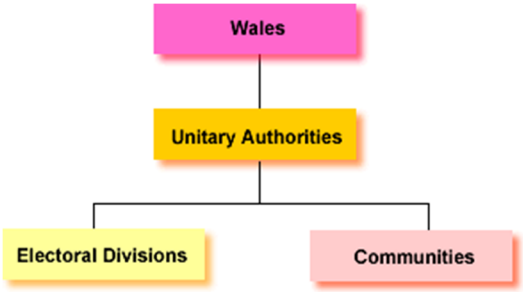
Note however that the diagram shows the geographic structure rather than the administrative reporting structure.

In practice, neither metropolitan counties nor Regions (former Government Office Regions) are truly part of the administrative

hierarchy, and electoral wards/divisions are simply the 'building blocks' from which higher units are constituted.

Parishes on the other hand can have their own council, but have been isolated from the geographic structure as, unlike electoral wards/divisions, they are not found across the whole of England.

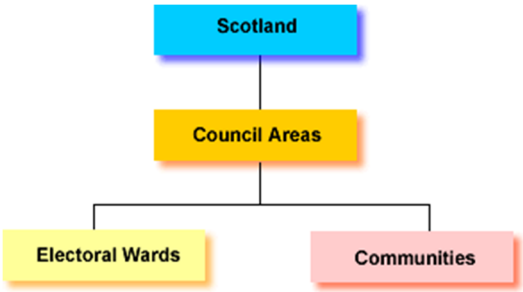
Wales



Wales is subject to the administration of both the UK Government in Westminster and also the National Assembly for Wales in Cardiff.

Wales is subdivided into 22 unitary authorities, which in turn are divided into electoral divisions and communities.

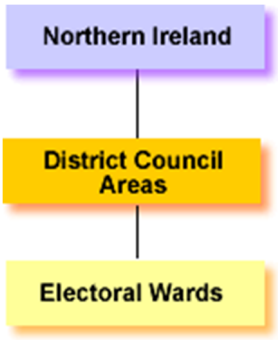
Scotland



Scotland is subject to the administration of both the UK Government in Westminster and also the Scottish Government in Edinburgh.

Scotland is subdivided into 32 council areas, which in turn are divided into electoral wards and communities.

Northern Ireland



Northern Ireland is subject to the administration of both the UK Government in Westminster and also the Northern Ireland Executive in Belfast.

Northern Ireland is subdivided into 26 district council areas (although within Northern Ireland they are also known as 'local government districts', which in turn are divided into electoral wards.

Census Geography
In the context of this user engagement exercise 'census geography' refers to the base unit for census data releases, namely Output Areas (OAs). In addition Lower Layer Super Output Areas (LSOAs) in England & Wales, Data Zones in Scotland, Super Output Areas (SOAs) in Northern Ireland and Middle Layer Super Output Areas (MSOAs) are considered to be part of census geography as they are formed using Output Areas.

County
Please see 'Administrative Geographies' which details the hierarchical nature of different levels of local government in England.

Data source
In the context of this consultation, 'data source' refers to an online location where multiple datasets can be accessed and downloaded.

Data Zones
The Scottish equivalent of Lower Layer Super Output Areas in England and Wales. Constructed using Output Areas there are 6,505 covering Scotland with an average resident population of 750.

Dataset
In the context of this consultation, 'dataset' refers to a single table of data. An example would be a key statistic table from the 2001 UK Census.

Geodemographic classification
The description of people according to where they live derived from the study of spatial information.

Index of Multiple Deprivation
Separate country Indexes of Multiple Deprivation measures relative deprivation across the UK. They separately combine a number of the same or similar indicators, chosen to cover a range of economic, social and housing issues, into a single deprivation score for each small area in England, Wales (for LSOAs), Scotland (for Data Zones) and Northern Ireland (for SOAs). This allows each area to be ranked relative to one another in each country according to their level of deprivation.

Local Authority
In England there are five different types of local authority: metropolitan, unitary, London boroughs, county councils and district councils. These different types are all included under the umbrella term 'local authority'. Please see 'Administrative Geographies', which details how these different types of local authority fit into the hierarchical nature of local government in England. These divisions and names can

differ in Wales, Scotland and Northern Ireland, but follow a similar hierarchical structure.

Lower Layer Super Output Areas (LSOAs)
Constructed using Output Areas. There are 34,378 in England and Wales with an average population of 1,500 and an average number of households of 400 (as at 2001).

Middle Layer Super Output Areas (MSOAs)
Constructed using Output Areas. There are 7,193 in England and Wales with an average population of 7,200 and an average number of households of 2,000 (as at 2001).

Non-Census data
In the context of this consultation form 'non-census data' refers to all available data (normally freely available open data) that are not derived from any UK census sources. An example would be administrative data from central or local government.

Open Data
In the context of this consultation, 'open data' refers to data that are freely available (although may require a fee to access). They may supplement or substitute for (freely available) census data. 'Open data' usually derive from any of a range of government agencies, and are provided in the interests of creating more accountable, transparent, participatory and collaborative government.

Output Areas (OAs)
The smallest dissemination units available for census data. They were designed to have similar population sizes and to be as socially homogenous as possible. Based on 2001 Census data and postcodes in use in 2000-2001, there are 223,060 Output Areas covering the UK with an average population of 264 and an average number of households of 110 (as at 2001).

Regional
Please see 'Administrative Geographies' which details the hierarchical nature of different levels of local government in England.

Super Output Areas (SOAs)
In the context of this consultation SOAs are referring to the areal units created by the Northern Ireland Statistics & Research Agency (NISRA). Like LSOAs in England and Wales they are constructed using Output Areas with 890 covering Northern Ireland with an average population of 2,000. SOAs were also created for England and Wales, but due to the larger population are sub-divided into LSOAs and MSOAs that are referred to in this consultation form.

Wards
A constituent part of an electoral district that is the primary unit of UK administrative and electoral geography.

The differences between **Output Areas (OAs), Lower Layer Super Output Areas (LSOAs)** and **Middle Layer Super Output Areas (MSOAs)**

Clicking on the links above will display maps are using the population of White British in London in 2001 to illustrate the differences in using OA, LSOA and MSOA output geography.

To give you an idea what the differences are in numbers; in London there are currently 24,140 OAs, 4,765 LSOAs and 983 MSOAs based on postcodes in use on the day of the 2001 UK Census.

These numbers are likely to change slightly as they are updated for use with the 2011 Census data outputs. To put this change into context nationally, less than 5% of the current 223,060 Output Areas will be modified. This means at least roughly 212,000 Output Areas will remain consistent across the UK.

Appendix B

B.1. The 2011 Area Classification for Output Areas User Engagement

Appendix B provides full tabular results of the 2011 OAC user engagement and a selection of general comments made by respondents and those specifically related to the 2011 OAC. The results are presented as percentages with counts in brackets unless otherwise stated.

B.1.1. Findings from the User Engagement

Table B.1: Responses by stakeholder group

Respondent Type	Responses	Percentage of total
Local Authorities (LA)	19	50
Central Government (CG)	2	5
Health (H)	3	8
Other Public Sector (PS)	3	8
Commercial Organisations & Individuals (CO)	7	18
Academia (A)	4	11
Total (All)	**38**	**100**

B.1.1.1. The current 2001 Area Classification for Output Areas

Question 1: Do you know what the current 2001 Area Classification for Output Areas (2001 OAC) is?

Table B.1: Responses to Question 1 of the 2011 OAC user engagement

	All	LA	CG	H	PS	CO	A
Yes	97 (37)	100 (19)	50 (1)	100 (3)	100 (3)	100 (7)	100 (4)
No	3 (1)	0 (0)	50 (1)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

Question 2: Do you (or your organisation) currently use the 2001 OAC?

Table B.2: Responses to Question 2 of the 2011 OAC user engagement

	All	LA	CG	H	PS	CO	A
Yes	61 (23)	47 (19)	0 (1)	100 (3)	67 (3)	71 (7)	100 (4)
No	39 (15)	53 (0)	100 (1)	0 (0)	33 (0)	29 (0)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

Question 2a: If you answered "Yes" how long have you (or your organisation) been using the 2001 OAC for?

Table B.3: Responses to Question 2a of the 2011 OAC user engagement

	All	LA	CG	H	PS	CO	A
1 year or less	13 (3)	11 (1)	0 (0)	0 (0)	0 (0)	20 (1)	25 (1)
2 to 5 years	39 (9)	44 (4)	0 (0)	0 (0)	100 (2)	20 (1)	50 (2)
Over 5 years	48 (11)	44 (4)	0 (0)	100 (3)	0 (0)	60 (3)	25 (1)
Total	**100 (23)**	**100 (9)**	**0 (0)**	**100 (3)**	**100 (2)**	**100 (5)**	**100 (4)**

Question 3: If you answered "No" to Question 2 have you (or your organisation) previously used the 2001 OAC?

Table B.4: Responses to Question 3 of the 2011 OAC user engagement

	All	LA	CG	H	PS	CO	A
Yes	67 (10)	60 (6)	50 (1)	0 (0)	100 (1)	100 (2)	0 (0)
No	27 (4)	30 (3)	50 (1)	0 (0)	0 (0)	0 (0)	0 (0)
No Answer	7 (1)	10 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (15)**	**100 (10)**	**100 (2)**	**0 (0)**	**100 (1)**	**100 (2)**	**0 (0)**

3a: If you answered "No" why have you never used the 2001 OAC?

Table B.5: Responses to Question 3a of the 2011 OAC user engagement

	All	LA	CG	H	PS	CO	A
Not useful	67 (2)	50 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Did not know it existed	27 (1)	0 (0)	100 (1)	0 (0)	0 (0)	0 (0)	0 (0)
No Answer	7 (2)	50 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (5)**	**100 (4)**	**100 (1)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**

3b: If you answered "Yes" how long ago did you (or your organisation) stop using the 2001 OAC?

Table B.6: Responses to Question 3b of the 2011 OAC user engagement

	All	LA	CG	H	PS	CO	A
1 year or less	18 (2)	14 (1)	0 (0)	0 (0)	0 (0)	50 (1)	0 (0)
2 to 5 years	55 (6)	43 (3)	100 (1)	0 (0)	100 (1)	50 (1)	0 (0)
Over 5 years	18 (2)	29 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
No Answer	9 (1)	14 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (11)**	**100 (7)**	**100 (1)**	**0 (0)**	**100 (1)**	**100 (2)**	**0 (0)**

3c: If you answered "Yes" why did you stop using the 2001 OAC?

Table B.7: Responses to Question 3c of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
Use Commercial Systems instead	42 (5)	38 (3)	0 (0)	0 (0)	100 (1)	50 (1)	0 (0)
Need finer granularity than OA	17 (2)	25 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Not realistic	8 (1)	13 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Not used	17 (2)	13 (1)	0 (0)	0 (0)	0 (0)	50 (1)	0 (0)
Could not use with non-OA datasets	8 (1)	0 (0)	100 (1)	0 (0)	0 (0)	0 (0)	0 (0)
No Answer	8 (1)	13 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (12)**	**100 (8)**	**100 (1)**	**0 (0)**	**100 (1)**	**100 (1)**	**0 (0)**

Question 4: What alternative commercial geodemographic classifications do you (or your organisation) use?

Table B.8: Responses to Question 4 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
ACORN (CACI)	31 (14)	30 (6)	100 (2)	0 (0)	25 (1)	27 (3)	40 (2)
Mosaic (Experian)	31 (14)	40 (8)	0 (0)	67 (2)	25 (1)	27 (3)	0 (0)
P^2 People & Places (Beacon Dodsworth)	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	20 (1)
PersonicX (Acxiom)	4 (2)	5 (1)	0 (0)	0 (0)	25 (1)	0 (0)	0 (0)
CallCredit's tools	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)	9 (1)	0 (0)
Audiences Insight Profiles	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)	9 (1)	0 (0)
None	22 (10)	20 (4)	0 (0)	0 (0)	25 (1)	27 (3)	40 (2)
No Answer	4 (2)	5 (1)	0 (0)	33 (1)	0 (0)	0 (0)	0 (0)
Total	**100 (45)**	**100 (20)**	**100 (2)**	**100 (3)**	**100 (4)**	**100 (11)**	**100 (5)**

Question 5: Please indicate the geographical coverage(s) you favour when using a geodemographic classification?

Table B.9: Responses to Question 5 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
UK	11 (14)	5 (3)	22 (2)	8 (1)	20 (2)	19 (4)	13 (2)
Country	11 (13)	9 (5)	22 (2)	17 (2)	10 (1)	5 (1)	13 (2)
Regional	9 (11)	9 (5)	11 (1)	8 (1)	10 (1)	10 (2)	6 (1)
County	11 (13)	13 (7)	11 (1)	8 (1)	10 (1)	10 (2)	6 (1)
Local Authority	24 (30)	29 (16)	11 (1)	17 (2)	20 (2)	24 (5)	25 (4)
City, Town or Village	15 (19)	16 (9)	11 (1)	17 (2)	20 (2)	14 (3)	13 (2)
District or Area of a City, Town or Village	18 (22)	16 (9)	11 (1)	25 (3)	10 (1)	19 (4)	25 (4)
No Answer	1 (1)	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (123)**	**100 (55)**	**100 (9)**	**100 (12)**	**100 (10)**	**100 (21)**	**101 (16)**

Question 6: Would you welcome a new version of the 2001 OAC?

Table B.10: Responses to Question 6 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
Yes	89 (34)	95 (18)	100 (2)	67 (2)	100 (3)	86 (6)	75 (3)
No	8 (3)	0 (0)	0 (0)	33 (1)	0 (0)	14 (1)	25 (1)
No Answer	3 (1)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

Question 7: Should a new 2011 Area Classification for Output Areas (2011 OAC) be a general purpose classification (like the current 2001 OAC), or should it focus on producing specialised variants (such as health, education or crime)?

Table B.11: Responses to Question 7 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
General purpose	55 (21)	53 (10)	100 (2)	33 (1)	33 (1)	57 (4)	75 (3)
Specialised variants	45 (17)	47 (9)	0 (0)	67 (2)	67 (2)	43 (3)	25 (1)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

Question 8: Flexibility in specifying the variables that are to make up the 2011 OAC would open up a range of options for area classification using Open Government Data. Is it important to you that the 2011 OAC be directly comparable – in terms of similar census data being used to construct it - with the 2001 OAC?

Table B.12: Responses to Question 8 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
Yes	13 (5)	11 (2)	0 (0)	67 (2)	0 (0)	0 (0)	25 (1)
No	87 (33)	89 (17)	100 (2)	33 (1)	100 (3)	100 (7)	75 (3)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

Question 8a: If you answered 'No' then what are the other priorities that are important to you in the construction of the 2011 OAC?

Table B.13: Responses to Question 8a of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
Updateable	36 (18)	36 (9)	50 (1)	50 (1)	50 (2)	36 (5)	0 (0)
Better variables	46 (23)	44 (11)	0 (0)	50 (1)	50 (2)	50 (7)	67 (2)
Relevant to specific work areas	2 (1)	4 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Better representation of local areas	6 (3)	12 (3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Use of Open Data sources	4 (2)	4 (1)	0 (0)	0 (0)	0 (0)	7 (1)	0 (0)
More documentation	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)	7 (1)	0 (0)
Creating a robust classification	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	33 (1)
No Answer	2 (1)	0 (0)	50 (1)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (50)**	**100 (25)**	**100 (2)**	**100 (2)**	**100 (4)**	**100 (14)**	**100 (3)**

Question 9: The 2001 OAC divides the population of the UK into 7 Supergroups, 21 Groups and 52 Subgroups. How would you describe this framework when using the 2001 OAC for your particular purposes?

Table B.14: Responses to Question 9 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
Extremely limited	3 (1)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Limited	15 (6)	15 (3)	0 (0)	67 (2)	33 (1)	0 (0)	0 (0)
Satisfactory	44 (17)	45 (9)	50 (1)	33 (1)	33 (1)	43 (3)	50 (2)
Good	31 (12)	30 (6)	50 (1)	0 (0)	33 (1)	29 (2)	50 (2)
Excellent	3 (1)	0 (0)	0 (0)	0 (0)	0 (0)	14 (1)	0 (0)
No Answer	5 (2)	5 (1)	0 (0)	0 (0)	0 (0)	14 (1)	0 (0)
Total	**101 (39)**	**100 (20)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

B.1.1.2. New for the 2011 Area Classification for Output Areas

Question 10: Thinking about how you use and interpret the 2001 OAC, how useful do you think each to the following options would be to you for the 2011 OAC? (1 = Not at all useful to 5 = Extremely useful)

10a. Maps in PDF (or similar) format that are not interactive

Table B.15: Responses to Question 10a of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
1	21 (8)	16 (3)	0 (0)	33 (1)	67 (2)	29 (2)	0 (0)
2	32 (12)	42 (8)	50 (1)	0 (0)	33 (1)	29 (2)	0 (0)
3	24 (9)	26 (5)	0 (0)	33 (1)	0 (0)	29 (2)	25 (1)
4	21 (8)	11 (2)	50 (1)	33 (1)	0 (0)	14 (1)	75 (3)
5	3 (1)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Do not know	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

10b. Online interactive maps with clickable details

Table B.16: Responses to Question 10b of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
1	5 (2)	11 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
2	3 (1)	0 (0)	0 (0)	0 (0)	0 (0)	14 (1)	0 (0)
3	16 (6)	16 (3)	50 (1)	0 (0)	0 (0)	29 (2)	0 (0)
4	42 (16)	42 (8)	50 (1)	67 (2)	33 (1)	43 (3)	25 (1)
5	34 (13)	32 (6)	0 (0)	33 (1)	67 (2)	14 (1)	75 (3)
Do not know	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

10c. Mapping against different backdrops (such as Google Maps or OpenStreetMap)

Table B.17: Responses to Question 10c of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
1	8 (3)	16 (3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
2	8 (3)	0 (0)	100 (2)	0 (0)	33 (1)	0 (0)	0 (0)
3	16 (6)	21 (4)	0 (0)	0 (0)	0 (0)	29 (2)	0 (0)
4	42 (16)	42 (8)	0 (0)	67 (2)	33 (1)	43 (3)	50 (2)
5	26 (10)	21 (4)	0 (0)	33 (1)	33 (1)	29 (2)	50 (2)
Do not know	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**99 (3)**	**101 (7)**	**100 (4)**

10d. Correlation tables (showing to what extent the variables within the classification correlate with each other)

Table B.18: Responses to Question 10d of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
2	5 (2)	11 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
3	11 (4)	5 (1)	50 (1)	33 (1)	0 (0)	0 (0)	25 (1)
4	42 (16)	47 (9)	50 (1)	67 (2)	0 (0)	43 (3)	25 (1)
5	37 (14)	26 (5)	0 (0)	0 (0)	100 (3)	57 (4)	50 (2)
Do not know	5 (2)	11 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

10e. Bar graphs of the group's attributes

Table B.19: Responses to Question 10e of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
1	5 (2)	5 (1)	0 (0)	0 (0)	33 (1)	0 (0)	0 (0)
2	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
3	21 (8)	16 (3)	50 (1)	33 (1)	33 (1)	14 (1)	25 (1)
4	50 (19)	47 (9)	50 (1)	67 (2)	33 (1)	43 (3)	75 (3)
5	24 (9)	32 (6)	0 (0)	0 (0)	0 (0)	43 (3)	0 (0)
Do not know	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

10f. Radial plots of the group's attributes

Table B.20: Responses to Question 10f of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
1	8 (3)	5 (1)	0 (0)	0 (0)	33 (1)	0 (0)	25 (1)
2	5 (2)	5 (1)	0 (0)	33 (1)	0 (0)	0 (0)	0 (0)
3	26 (10)	37 (7)	50 (1)	0 (0)	33 (1)	14 (1)	0 (0)
4	32 (12)	26 (5)	50 (1)	67 (2)	33 (1)	29 (2)	25 (1)
5	24 (9)	21 (4)	0 (0)	0 (0)	0 (0)	43 (3)	50 (2)
Do not know	5 (2)	5 (1)	0 (0)	0 (0)	0 (0)	14 (1)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

Question 11: Thinking about your own understanding of the existing 2001 OAC, how useful do you think each of the following options would be to you for the 2011 OAC? (1 = Not at all useful to 5 = Extremely useful)

11a. Group Name

Table B.21: Responses to Question 11a of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
2	3 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	25 (1)
3	5 (2)	0 (0)	0 (0)	33 (1)	0 (0)	0 (0)	25 (1)
4	34 (13)	42 (8)	50 (1)	33 (1)	0 (0)	29 (2)	25 (1)
5	53 (20)	53 (10)	50 (1)	33 (1)	67 (2)	71 (5)	25 (1)
Do not know	5 (2)	5 (1)	0 (0)	0 (0)	33 (1)	0 (0)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

11b. Graphical Representation (radial plots and bar graphs)

Table B.22: Responses to Question 11b of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
2	5 (2)	0 (0)	0 (0)	0 (0)	33 (1)	14 (1)	0 (0)
3	18 (7)	21 (4)	0 (0)	0 (0)	0 (0)	43 (3)	0 (0)
4	45 (17)	37 (7)	50 (1)	100 (3)	33 (1)	29 (2)	75 (3)
5	24 (9)	32 (6)	0 (0)	0 (0)	33 (1)	14 (1)	25 (1)
Do not know	8 (3)	11 (2)	50 (1)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (38)**	**101 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

11c. Group definitions (a written summary of the key characteristics of each group)

Table B.23: Responses to Question 11c of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
2	3 (1)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
3	3 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	25 (1)
4	34 (13)	32 (6)	50 (1)	33 (1)	0 (0)	29 (2)	75 (3)
5	58 (22)	63 (12)	50 (1)	67 (2)	67 (2)	71 (5)	0 (0)
Do not know	3 (1)	0 (0)	0 (0)	0 (0)	33 (1)	0 (0)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

11d. Key points of characteristics you would expect to find in each group

Table B.24: Responses to Question 11d of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
2	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
3	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
4	32 (12)	16 (3)	100 (2)	33 (1)	0 (0)	29 (2)	100 (4)
5	66 (25)	84 (16)	0 (0)	67 (2)	67 (2)	71 (5)	0 (0)
Do not know	3 (1)	0 (0)	0 (0)	0 (0)	33 (1)	0 (0)	0 (0)
Total	**101 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

11e. Written 'pen portraits' of typical households found within each group

Table B.25: Responses to Question 11e of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
2	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
3	13 (5)	5 (1)	0 (0)	0 (0)	0 (0)	14 (1)	75 (3)
4	29 (11)	26 (5)	100 (2)	33 (1)	33 (1)	14 (1)	25 (1)
5	58 (22)	68 (13)	0 (0)	67 (2)	67 (2)	71 (5)	0 (0)
Do not know	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**101 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

11f. Written 'pen portraits' of typical housing and built environments found in each group

Table B.26: Responses to Question 11f of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
2	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
3	18 (7)	11 (2)	50 (1)	0 (0)	0 (0)	14 (1)	75 (3)
4	32 (12)	32 (6)	0 (0)	33 (1)	33 (1)	43 (3)	25 (1)
5	47 (18)	58 (11)	0 (0)	67 (2)	67 (2)	43 (3)	0 (0)
Do not know	3 (1)	0 (0)	50 (1)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**101 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

Question 12: Do you agree with the view that it would be helpful to adjust the composition of each group for different parts of the UK (so, for example, there might be separate classifications made for London, or Scotland)?

Table B.27: Responses to Question 12 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
Strongly disagree	13 (5)	11 (2)	0 (0)	0 (0)	33 (1)	14 (1)	25 (1)
Disagree	26 (10)	21 (4)	50 (1)	33 (1)	33 (1)	14 (1)	50 (2)
Neither agree nor disagree	18 (7)	16 (3)	50 (1)	33 (1)	0 (0)	14 (1)	25 (1)
Agree	37 (14)	42 (8)	0 (0)	33 (1)	33 (1)	57 (4)	0 (0)
Strongly agree	5 (2)	11 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

Question 13: Please identify what, if any, extra features would you like the 2011 OAC to have when compared with the 2001 OAC

Table B.28: Responses to Question 13 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
OA Code Alias	2 (1)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
OAC Profiler	5 (2)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	20 (1)
Pen Portraits	17 (7)	20 (4)	50 (1)	0 (0)	0 (0)	25 (2)	0 (0)
Diversity indicator	2 (1)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Updateable variables	2 (1)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Linkages to other data	10 (4)	5 (1)	0 (0)	0 (0)	33 (1)	13 (1)	20 (1)
Names for all groups	10 (4)	5 (1)	0 (0)	0 (0)	0 (0)	25 (2)	20 (1)
Commercial features	5 (2)	0 (0)	0 (0)	33 (1)	0 (0)	13 (1)	0 (0)
Raised awareness	2 (1)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
No comment	44 (18)	45 (9)	50 (1)	67 (2)	67 (2)	25 (2)	40 (2)
Total	**100 (41)**	**100 (20)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (8)**	**100 (5)**

B.1.1.3. Dissemination of the 2011 Area Classification for Output Areas

Question 14: Which methods of dissemination for the 2011 OAC would you be most likely to use?

Table B.29: Responses to Question 14 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
Online interactive mapping	15 (19)	15 (10)	0 (0)	23 (3)	18 (2)	9 (2)	17 (2)
Enhanced online interactive mapping	18 (23)	19 (13)	0 (0)	23 (3)	27 (3)	9 (2)	17 (2)
Microsoft Excel/CSV file(s)	27 (35)	27 (18)	33 (2)	15 (2)	18 (2)	32 (7)	33 (4)
Software to append the 2011 OAC codes to postcodes	18 (24)	15 (10)	33 (2)	23 (3)	9 (1)	27 (6)	17 (2)
Digital Boundary Data	23 (30)	24 (16)	33 (2)	15 (2)	27 (3)	23 (5)	17 (2)
Total	**100 (131)**	**100 (67)**	**100 (6)**	**100 (13)**	**100 (11)**	**100 (22)**	**100 (12)**

Question 15: Other data sources could be used to give greater context to the 2011 OAC. Rather than contributing to the classification itself, these could be used to help visualise the 2011 OAC in different ways. What (if any) data sources would you like to be able to use alongside the final 2011 OAC output?

Table B.30: Responses to Question 15 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
Index of Multiple Deprivation (IMD)	62 (33)	75 (18)	100 (2)	60 (3)	60 (3)	50 (5)	29 (2)
Temporal data	15 (8)	13 (3)	0 (0)	0 (0)	20 (1)	30 (3)	14 (1)
Health related data	4 (2)	0 (0)	0 (0)	40 (2)	0 (0)	0 (0)	0 (0)
Land use data	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	14 (1)
Weather history data	2 (1)	0 (0)	0 (0)	0 (0)	20 (1)	0 (0)	0 (0)
Travel to Work areas	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	14 (1)
Other survey data	11 (6)	8 (2)	0 (0)	0 (0)	0 (0)	20 (2)	28 (2)
No Answer	2 (1)	4 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (53)**	**100 (24)**	**100 (2)**	**100 (5)**	**100 (5)**	**100 (10)**	**100 (7)**

B.1.1.4. Construction of the 2011 Area Classification for Output Areas

Question 16: There are multiple levels of spatial resolution that data can be produced. In addition to Output Areas are there any other spatial resolutions you believe would benefit from having their own classification?

Table B.31: Responses to Question 16 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
LSOA/DZ/SOA	30 (32)	34 (16)	25 (2)	14 (1)	27 (3)	29 (6)	36 (4)
MSOA	10 (11)	9 (4)	13 (1)	14 (1)	9 (1)	14 (3)	9 (1)
USOA	1 (1)	0 (0)	0 (0)	0 (0)	9 (1)	0 (0)	0 (0)
Postcode	3 (3)	0 (0)	0 (0)	0 (0)	0 (0)	14 (3)	0 (0)
Ward	22 (23)	28 (13)	13 (1)	0 (0)	18 (2)	19 (4)	27 (3)
Parishes	2 (2)	4 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Parliament Constituency	1 (1)	0 (0)	0 (0)	0 (0)	9 (1)	0 (0)	0 (0)
Local Authorities	17 (18)	17 (8)	25 (2)	14 (1)	9 (1)	19 (4)	18 (2)
Counties	5 (5)	4 (2)	13 (1)	14 (1)	9 (1)	0 (0)	0 (0)
Regions	7 (7)	4 (2)	13 (1)	14 (1)	9 (1)	5 (1)	9 (1)
No Answer	2 (2)	0 (0)	0 (0)	29 (2)	0 (0)	0 (0)	0 (0)
Total	**100 (105)**	**100 (47)**	**100 (8)**	**100 (7)**	**100 (11)**	**100 (21)**	**100 (11)**

Question 17: The 2001 OAC uses only 2001 Census data in its construction. It has been suggested that, in addition to using 2011 Census data, it might be possible for the 2011 OAC to be enhanced with supplementary non-census open data sources, and updated periodically over time. Would you find this beneficial?

Table B.32: Responses to Question 17 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
Yes	87 (33)	95 (18)	0 (0)	100 (3)	100 (3)	100 (7)	50 (2)
No	5 (2)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	25 (1)
Do not know	8 (3)	0 (0)	100 (2)	0 (0)	0 (0)	0 (0)	25 (1)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

Question 18: It is unlikely that many open data sources will offer UK wide coverage. What extent of coverage do you believe is a minimum requirement for an acceptable general purpose classification?

Table B.33: Responses to Question 18 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
UK only	10 (4)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	75 (3)
Countrywide	43 (18)	43 (9)	50 (1)	0 (0)	40 (2)	71 (5)	25 (1)
Regional	17 (7)	19 (4)	0 (0)	0 (0)	20 (1)	29 (2)	0 (0)
Local Authority	26 (11)	33 (7)	0 (0)	100 (3)	20 (1)	0 (0)	0 (0)
Ward	2 (1)	0 (0)	0 (0)	0 (0)	20 (1)	0 (0)	0 (0)
No Answer	2 (1)	0 (0)	50 (1)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (42)**	**100 (21)**	**100 (2)**	**100 (3)**	**100 (5)**	**100 (7)**	**100 (4)**

Question 19: If the 2011 OAC could be updated with new data, how frequently should this be done?

Table B.34: Responses to Question 19 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
Once a year	42 (16)	37 (7)	0 (0)	67 (2)	33 (1)	71 (5)	25 (1)
Every two years	26 (10)	37 (7)	50 (1)	33 (1)	33 (1)	0 (0)	0 (0)
Every three years or longer	29 (11)	26 (5)	50 (1)	0 (0)	33 (1)	29 (2)	50 (2)
No Answer	3 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	25 (1)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

Question 20: Change in the social, economic and demographic structure of areas in the UK occurs at different rates. Instead of updating the 2011 OAC it might be possible to use non-census sources to flag areas where population changes have occurred, enabling the user to recognise parts of the UK where the classification had probably become unreliable. Would you find this helpful?

Table B.35: Responses to Question 20 of the 2011 OAC user engagement

	All	**LA**	**CG**	**H**	**PS**	**CO**	**A**
Yes	97 (37)	95 (18)	100 (2)	100 (3)	100 (3)	100 (7)	100 (4)
No	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Do not know	3 (1)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	**100 (38)**	**100 (19)**	**100 (2)**	**100 (3)**	**100 (3)**	**100 (7)**	**100 (4)**

B.1.1.5. General Comments

- I think the [2001] OAC is a brilliant resource. [respondent 5]
- While I support the improvement of any Government data, I do think a careful assessment of the economic damage that could be caused to commercial products such as Acorn or Mosaic. While the availability of more data is a good thing, this should be done in conjunction with the private sector; not in competition with it. [respondent 28]
- An incredibly valuable resource that in my view could be made more so. Commercial systems are very costly and an open source classification that helps organisations understand more about the communities they serve is invaluable. [respondent 31]
- I am wondering if it might be easier to use Lower Super Output Areas for the base unit for the new classification instead of Output Areas. I am thinking that there may be more other Open Data sources available at this level than at the Output Area level. [respondent 32]
- A classification which could encompass population change 1991-2011 as well as 2011 characteristics and which could be applied to all three Censuses would be very useful. [respondent 35]
- [The 2001] OAC is a great classification, and I really appreciate its transparency, basis in open source data, and commitment to ground-truthing and academic use and review. In many ways, I feel this makes it superior to all the commercially available classifications around, and I'd be sad to see it move away from this model (although I do understand that getting it more widely used is probably a priority). I'm a little hesitant about pithy group names and pen portraits for this reason, although I can see how some other users at my institution might find such representations the OA characteristics more accessible. [respondent 38]

B.1.1.6. Comments regarding the 2011 Area Classification for Output Areas

- We consider that a general-purpose classification is an essential output. However, specialised variant classifications covering health, education and crime are becoming more relevant. [respondent 1]
- [The] 2011 OAC should from the outset offer a credible alternative to commercial classifications with the opportunities for significant cost savings across the public sector organisations. [respondent 4]

- I would like to see [the 2011] OAC developed to the extent that it could be used in place of commercial data. This would reduce the cost of customer insight and encourage more Local Authorities to develop a greater understanding of the people living in their communities. [respondent 8]
- We would prefer the 2011 OAC to be produced more timely than the 2001 OAC. [respondent 16]
- It may be helpful if the 2001 OAC could be reproduced based on the new methodology to be consistent/compatible with the new 2011 OAC. [respondent 20]
- Outputs should include groups of 200 / 2,000 database analysis [respondent 33]
- Subsidiary specialist outputs made available - for example distance of Output Area from cluster centroid, an advantage of the 2011 OAC being transparent [respondent 33]
- Marketing of the 2011 OAC is vital. [respondent 33]
- Rehearse the creation of 2011 OAC by using 2001 Census data. [respondent 33]
- Name the clusters: this might be done informally on a mirror site, such as the OAC User Group, rather than putting ONS in a difficult situation. [respondent 33]
- Create a user guide, giving simple descriptions of the clusters. [respondent 33]
- Code some surveys, such as the British Population Survey, to help with cluster descriptions. [respondent 33]
- Make it easy for users to download not only [the 2011] OAC, but also the Coder (to code postcode files), ONS's Output Area boundaries and background mapping. [respondent 33]
- ONS should plan to rapidly add the 2011 OAC to its sample surveys such as the Living Costs, Food Survey and the Wealth Assets Survey. This will create much additional value. [respondent 33]
- The unique selling point for the OAC is an open and reproducible methodology - any deviation from this will be detrimental to the overarching aims of the classification. [respondent 37]

Appendix C

C.1. Pen Portraits for the 2011 OAC

The following pen portraits provide descriptions of the likely dominant characteristics of the residents who live in areas assigned to one of the 8 Supergroups, 26 Groups and 76 Subgroups of the 2011 OAC. The descriptions for the Groups and Subgroups are organised under their parent Supergroup heading.

C.1.1. Rural Residents

1 – Rural Residents

The population of this Supergroup live in rural areas that are far less densely populated compared with elsewhere in the country. They will tend to live in large detached properties which they own and work in the agriculture, forestry and fishing industries. The level of unemployment in these areas is below the national average. Each household is likely to have multiple motor vehicles, and these will be the preferred method of transport to their places of work. The population tends to be older, married and well educated. An above average proportion of the population in these areas provide unpaid care and an above average number of people live in communal establishments (most likely to be retirement homes). There is less ethnic integration in these areas and households tend to speak English or Welsh as their main language.

1a – Farming Communities

This group has a lower population density than the Supergroup, and people are less likely to live in communal establishments. There is a higher proportion of households living in detached properties when compared with the Supergroup and much lower proportions living in terraced properties and flats. Households are less likely to live in social rented accommodation. There is a higher proportion of people working in the agriculture industry compared with the Supergroup.

1a1 – Rural Workers and Families

The population of this Subgroup has a slightly higher proportion of people aged 0 to 4 when compared with the Group. Households are slightly more likely to live in detached properties, and less likely to live in other types of property. They are also slightly more likely to live in overcrowded conditions than the Group and for household members to be unemployed.

1a2 – Established Farming Communities

Compared with the parent Supergroup, households in this Subgroup are more likely to live in a terraced or end-terraced house. People are slightly more likely to use public transport to get to work. Those in employment are more likely to work in financial related industries.

1a3 – Agricultural Communities

This Subgroup has a lower population density than the Group. Compared with the Group, a higher proportion of households live in terraced properties or flats, and privately rent their home. The proportion of people working in agricultural industries is higher than the parent Group.

1a4 – Older Farming Communities

The age make up of this Subgroup is lower than the Group for younger ages, but higher for ages 65 and over, and residents tend to live in more densely populated areas. Households are more likely to live in flats, though less likely to live in privately or socially rented accommodation. The proportion of people working in agricultural industries is lower than for the parent Group.

1b – Rural Tenants

The age structure is very similar to the Supergroup, though people are less likely to live in communal establishments. Compared with the parent Supergroup, there is a higher proportion of households living in semi-detached, terraced properties and flats, with a higher proportion socially renting. People are less likely to work in the agriculture industry than for the parent Supergroup.

1b1 – Rural Life

This Subgroup is slightly more densely populated than the parent Group. Households are less likely to live in flats. There are slightly more people working in the manufacturing and energy industries.

1b2 – Rural White-Collar Workers

This Subgroup is slightly less densely populated than the parent Group. When compared with the parent Group, a higher proportion of people work in the information and communication, and financial related industries, whilst unemployment is lower.

1b3 – Ageing Rural Flat Tenants

When compared with the parent Group there is a higher proportion of people who are aged 65 and over, and they live in slightly denser populated areas. A higher proportion of households live in flats and socially rent, whilst a lower proportion of people work in the information and communication, and financial industries.

1c – Ageing Rural Dwellers

The age structure of this group shows has a lower proportion of people aged under 65, and higher proportions aged 65 and over, particularly for the 90 and over age group. People are more likely to live in communal establishments or in detached properties.

1c1 – Rural Employment and Retirees

This Subgroup has a lower proportion of people aged 90 or over compared with the parent Group. It has a slightly higher proportion of people who were born in the EU and whose main language is not English or Welsh. The proportion of people working in agricultural industries is higher than the parent Group, and a higher proportion of households live in private rented accommodation.

1c2 – Renting Rural Retirement

This Subgroup has a higher proportion of people aged 90 and over compared with the parent Group. There is a higher proportion of households who live in terraced properties, and households are more likely to rent socially. The proportion of people working in agricultural industries is lower than the parent Group.

1c3 – Detached Rural Retirement

This Subgroup has a higher proportion of people aged 90 and over compared with the parent Group, a slightly higher proportion of households who live in detached properties, and a lower proportion who rent socially.

C.1.2. Cosmopolitans

2 – Cosmopolitans

The majority of the population in this Supergroup live in densely populated urban areas. They are more likely to live in flats and communal establishments, and private renting is more prevalent than nationally. The group has a high ethnic integration, with an above average number of residents from EU accession countries coinciding with a below average proportion of persons stating their country of birth as the UK or Ireland. A result of this is that households are less likely to speak English or Welsh as their main language. The population of the group is characterised by young adults, with a higher proportion of single adults and households without children than nationally. There are also higher proportions of full-time students. Workers are more likely to be employed in the accommodation, information and communication, and financial related industries, and using public transport, or walking or cycling to get to work.

2a – Students Around Campus

Compared with the parent Supergroup a higher proportion of people live in communal establishments. A lower proportion of people are married or divorced and a higher proportion are schoolchildren and full-time students. Households are more likely to live in terraced properties and to live in social rented accommodation compared with the national average. There is also a higher prevalence of workers in the accommodation or food service activities industries.

2a1 – Student Communal Living

This Subgroup has a high proportion of people (largely students) living in communal establishments compared with the Group. It also has a higher proportion of people who are of Chinese ethnicity. The proportion of people who are schoolchildren or full-time students is higher than the Group.

2a2 – Student Digs

When compared with the parent Group, this Subgroup has a much lower proportion of people living in communal establishments. There are lower proportions of people who are married or separated. The proportion of households with full-time students is higher than the parent Group, and households are more likely to be living in terraced houses and rented accommodation.

2a3 – Students and Professionals

The population in this Subgroup contains higher proportions of children aged 0 to 14, and adults aged 25 and over than the parent Group. The Subgroup has a lower proportion of people living in communal establishments than the parent Group, with higher proportions of people who are married or separated.

2b – Inner City Students

The age profile of this group shows a high proportion of schoolchildren, full-time students, and people aged 25 to 44, though a lower proportion married or divorced. Households are more likely to live in flats, to live in private rented accommodation, and to have overcrowded conditions. A lower proportion of people provide unpaid care, and a higher proportion work in accommodation or food service activities industries.

2b1 – Students and Commuters

The proportion of people who are white is slightly higher than for the parent Group, however the representation of all other ethnic groups is lower. The proportion of people with level 1 or 2, or apprenticeship qualifications is higher when compared with the parent Group. People are more likely to use private transport to travel to work.

2b2 – Multicultural Student Neighbourhood

The population in this Subgroup has a lower proportion of people aged 45 to 89 when compared with the parent Group. Its ethnic makeup has a higher proportion of persons of mixed ethnicity.

2c – Comfortable Cosmopolitans

The age profile of this group shows a higher proportion of people age 45 and over than the parent Supergroup. A higher proportion of people are divorced. There is a lower representation for all non-White ethnic groups when compared with the Supergroup and a lower proportion of people born in the old EU. There is a lower proportion of

households with full-time students and a higher proportion who live in flats. A higher proportion of workers are employed in the mining and manufacturing industries, and travel to work using private transport.

2c1 – Migrant Families

This Subgroup has a higher proportion of people aged 0 to 14 when compared with the parent Subgroup, with a higher proportion of residents of mixed ethnicity. Households are more likely to live in a detached, semi-detached or terraced property than the parent Group. A higher proportion of workers are employed in manufacturing industries.

2c2 – Migrant Commuters

The population in this Subgroup has a higher proportion of people who are of Pakistani ethnicity when compared with the parent Group, and households are more likely to live in socially rented accommodation. A lower proportion of households had two or more cars.

2c3 – Professional Service Cosmopolitans

This Subgroup had a lower proportion of people whose country of birth is in the new EU, and a lower proportion whose main language is not English or Welsh. When compared with the parent Group they are more likely to own their home, and less likely to live in overcrowded conditions. When compared with the other Subgroups for the parent Group, this Subgroup has the lowest proportion of people who are unemployed.

2d – Aspiring and Affluent

The proportion of people age 0 to 14 is higher than for the parent Supergroup. A higher proportion of people are married. There is a higher proportion of people who are of mixed ethnicity. A lower proportion of households have full-time students. Compared with the Supergroup a higher proportion of households live in semi-detached or terraced properties. People are more likely to work in the information and communication, and financial related industries, and use public transport to get to work.

2d1 – Urban Cultural Mix

When compared with the Group a higher proportion of people are of Indian ethnicity. A lower proportion of people were born in the old EU whereas a higher proportion were born in the new EU. They are more likely to live in a detached or semi-detached property.

2d2 – Highly-Qualified Quaternary Workers

The label 'quaternary' refers loosely to 'intellectual' activities. In comparison with the parent Group there is a higher proportion of people aged 5 to 14, and a lower proportion of persons of Indian ethnicity. Households are more likely to live in terraced or end-terraced properties.

2d3 – EU White-Collar Workers

A key characteristic for this Subgroup is the higher proportion of persons born in other EU countries, and most noticeably in other old EU countries, relative to the overall UK figure, and the parent Group. When compared with the parent Group, a higher proportion of the people in this Subgroup are Arab or of 'other' ethnic group. Households are more likely to live in flats and to be living in overcrowded conditions. Households are less likely to have two or more cars and also less likely to use private transport to travel to work.

C.1.3. Ethnicity Central

3 – Ethnicity Central

The population of this group is predominately located in the denser central areas of London, with other inner urban areas across the UK having smaller concentrations. All non-white ethnic groups have a higher representation than the UK average especially people of mixed ethnicity or who are Black, with an above average number of residents born in other EU countries. Residents are more likely to be young adults with slightly higher rates of divorce or separation than the national average, with a lower proportion of households having no children or non-dependent children. Residents are more likely to live in flats and more likely to rent. A higher proportion of people use public transport to get to work, with lower car ownership, and higher unemployment. Those in employment are more likely to work in the accommodation, information and communication, financial, and administrative related industries.

3a – Ethnic Family Life

When compared with the parent Supergroup, this group has a higher level of all non-White ethnic groups. There is a lower proportion of people born in the old EU but a higher proportion were born in the new EU. There is a higher proportion of people

whose main language is not English or Welsh. Households are more likely to live in detached, semi-detached or terraced properties.

3a1 – Established Renting Families

This Subgroup has a lower proportion of people who have Indian or Pakistani ethnicity when compared with the Group. The population is less likely to have been born in the new EU, and more likely to have dependent children. Households are more likely to be in socially rented accommodation.

3a2 – Young Families and Students

In comparison with the parent Group this Subgroup has a higher proportion of people who are of Indian or Pakistani ethnicity. Country of birth for residents is more likely to have been in one of the new EU countries, and residents are more likely to have higher level qualifications. Households are more likely to be privately renting than the parent Group and less likely to be social renting.

3b - Endeavouring Ethnic Mix

This group has a higher proportion of people who belong to the Bangladeshi ethnic group than the parent Supergroup but a lower proportion of those in Pakistani and Indian ethnic groups. There is a higher proportion of people who were born in the old EU countries. Households are more likely to live in flats and to socially rent than for the Supergroup. Overcrowding is also more prevalent, and public transport more commonly used to get to work.

3b1 – Striving Service Workers

The population of this Subgroup has a lower proportion of people who have Bangladeshi ethnicity than the Group but a higher proportion who have Black ethnicity. A lower proportion of households are likely to live in privately rented accommodation. Most of the other characteristics are similar to the parent Group.

3b2 – Bangladeshi Mixed Employment

A lower proportion of people in this Subgroup are of mixed or Black ethnic origin when compared with the Group but a far higher proportion of people have Bangladeshi ethnicity. A higher proportion of people whose main language is not English or Welsh are present in the Subgroup.

3b3 – Multi-Ethnic Professional Service Workers

When compared with the Group there is a higher proportion of people who are of Indian ethnicity but a lower proportion of Bangladeshi ethnicity. There is a higher proportion of people born in other EU countries with households more likely to live in privately rented accommodation in comparison with the parent Group.

3c – Ethnic Dynamics

In this group non-White ethnic groups are not represented as highly as in the parent Supergroup and there is a higher proportion of people born in the UK or Ireland. Households are more likely to live in a flat and to socially rent. There is a higher proportion of unemployed in the group but those in employment are more likely to work in the manufacturing industry, and to use private transport to travel to work.

3c1 – Constrained Neighbourhoods

In comparison with the Group, this Subgroup has a higher proportion of people who have mixed ethnicity. Households are more likely to live in terraced properties. People in work are slightly more likely to work in manufacturing industries, and households more likely to own two or more cars.

3c2 – Constrained Commuters

The population of this Subgroup has a lower proportion of people aged 65 and over than the parent Group. It also has a lower proportion of people with mixed ethnicity. Households in this Subgroup are more likely to live in flats, and to use public transport for getting to work.

3d – Aspirational Techies

With the exception of the Indian and mixed ethnic group, this group has a lower representation of all non-White ethnic groups than in the parent Supergroup. There is a higher proportion of people born in the old EU but a lower proportion whose main language is not English or Welsh. Households are more likely to live in semi-detached or terraced properties, and to live in privately rented accommodation. Workers are more likely to be employed in the information and communication industries, and to travel to work using public transport.

3d1 – New EU Tech Workers

The population of this Subgroup has a higher proportion of people who are of Indian or Pakistani ethnicity than the parent Group, and a higher proportion born in the new EU countries. Households are more likely to live in detached properties than the Group, and to live in privately rented accommodation. A higher proportion of people work in mining related industries, and use private transport for travelling to work.

3d2 – Established Tech Workers

The population of this Subgroup is slightly more likely to have Black ethnicity and more likely to be born in the UK or Ireland, and to have non-dependent children. There is a higher proportion of households who live in terraced housing, and a higher proportion living in socially rented accommodation. Households are also less likely to live in overcrowded conditions.

3d3 – Old EU Tech Workers

The population of this Subgroup is more likely to have Bangladeshi ethnicity, and to have been born in old EU countries. A higher proportion of households live in flats, and households are more likely to be living in socially rented accommodation.

C.1.4. Multicultural Metropolitans

4 – Multicultural Metropolitans

The population of this Supergroup is concentrated in larger urban conurbations in the transitional areas between urban centres and suburbia. They are likely to live in terraced housing that is rented – both private and social. The group has a high ethnic mix, but a below average number of UK and Irish born residents. A result of this is that households are less likely to speak English or Welsh as their main language. Residents are likely to be below retirement age. There is likely to be an above average number of families with children who attend school or college, or who are currently too young to do so. The rates of marriage and divorce are broadly comparable with the national average. The level of qualifications is just under the national average with the rates of unemployment being above the national average. Residents who are employed are more likely to work in the transport and administrative related industries. Public transport is the most likely method for individuals to get to and from work, since households are less likely to have multiple motor vehicles available to them.

4a – Rented Family Living

This group has a higher representation of White and mixed ethnicity residents than the Supergroup and a lower proportion of people whose main language is not English or Welsh. Households are more likely to live in terraced properties or flats, and to socially rent their property.

4a1 – Social Renting Young Families

This Subgroup, when compared with the parent Group, has a higher proportion of children aged 5 to 14, a higher proportion of people who have Pakistani ethnicity, and a higher proportion who were born in the UK or Ireland. Households are more likely to live in semi-detached properties, and to live in social rented accommodation. Unemployment is more prevalent when compared with the parent Group.

4a2 – Private Renting New Arrivals

When compared with the Group, this population of this Subgroup has a lower proportion of people who have Black or of mixed ethnicity. Residents are more likely to have been born in other EU countries. Households are more likely to be living in private rented accommodation.

4a3 – Commuters with Young Families

The population of this Subgroup has a lower proportion of people who are of Pakistani ethnicity, but a higher proportion of Black ethnicity when compared with the Group. Households are more likely to live in flats and to live in overcrowded conditions. People are more likely to work in the information and communication, and financial related industries.

4b – Challenged Asian Terraces

The population of this group has a higher proportion of non-White ethnic groups than the parent Supergroup especially people of the Pakistani ethnic group, and a higher proportion of 0 to 14 year-olds. It is more likely that their main language is not English or Welsh. A higher proportion of households live in terraced properties, and overcrowding is more prevalent. When compared with the Supergroup more people are likely to be unemployed, and those in employment to be working in the accommodation and food service industries.

4b1 – Asian Terraces and Flats

The population of this Subgroup has a higher representation of people from Indian, Black and Chinese ethnic groups, and a higher proportion of residents born in the new EU countries. Households are more likely to live in flats.

4b2 – Pakistani Communities

A key distinguishing feature of this Subgroup is the high proportion of people with Pakistani ethnicity, though residents are also more likely to have been born in the UK or Ireland, though less likely to speak English or Welsh as their main language. There is a slightly higher proportion of households who live in terraced housing, and also a higher proportion owning their property.

4c – Asian Traits

The population of this group has a higher proportion of people who are of Chinese ethnicity and particularly of Indian ethnicity. Compared with the parent Supergroup, households are more likely to live in detached and semi-detached properties, and to own their own home. A higher proportion of households have two or more cars, unemployment is lower, and workers are more likely to work in the Information and communication, and financial related industries.

4c1 – Achieving Minorities

The population of this Subgroup has a higher proportion of people who have Pakistani ethnicity, and lower proportions with Chinese and Black ethnicity than the parent Group. A lower proportion of residents were born in other EU countries. Households are more likely to live in detached and semi-detached properties, and to own their own property. Households are also less likely to live in overcrowded conditions.

4c2 – Multicultural New Arrivals

The population of this Subgroup has a higher representation of all non-White ethnic groups than the parent Group. A higher proportion of residents were born in new EU countries and a higher proportion without English or Welsh as their main language. Households are more likely to live in terraced properties or flats and to live in overcrowded conditions.

4c3 – Inner City Ethnic Mix

Compared with the parent Group, there is a higher representation of persons of mixed ethnicity, but lower representation for the other non-White ethnic groups, and a lower proportion of people whose main language is not English or Welsh. When compared with the parent Group, there is a lower proportion of households with non-dependent children, and households are more likely to live in flats.

C.1.5. Urbanites

5 – Urbanites

The population of this group are most likely to be located in urban areas in southern England and in less dense concentrations in large urban areas elsewhere in the UK. They are more likely to live in either flats or terraces, and to privately rent their home. The Supergroup has an average ethnic mix, with an above average number of residents from other EU countries. A result of this is households are less likely to speak English or Welsh as their main language. Those in employment are more likely to be working in the information and communication, financial, public administration and education related sectors. Compared with the UK, unemployment is lower.

5a – Urban Professionals and Families

The population of this group shows a noticeably higher proportion of children aged 0 to 14 than the parent Supergroup and a lower proportion aged 90 and over. There is also a higher proportion of people with mixed ethnicity. Households in this group are more likely to live in terraced properties and to live in privately rented accommodation. Unemployment is slightly higher than for the parent Supergroup.

5a1 – White Professionals

The population of this Subgroup has a lower representation of all ethnic groups, other than White when compared with the parent Group. Residents are less likely to have been born in other EU countries and more likely to have English or Welsh as their main language. Households are more likely to live in detached or semi-detached properties.

5a2 – Multi-Ethnic Professionals with Families

The population of this Subgroup has a higher representation of all non-White ethnic groups than the parent Group, and in particular representation of persons with Indian

or mixed ethnicity. Households are more likely to live in detached or semi-detached properties, and to live in socially rented accommodation. There is a higher proportion of people working in the information and communication, and financial related industries.

5a3 – Families in Terraces and Flats

When compared with the parent Group, this Subgroup has a higher proportion of households living in terraced properties or flats, with households more likely to rent their accommodation, either privately or socially. Households are more likely to live in overcrowded conditions than the parent Group and less likely to have two or more cars.

5b – Ageing Urban Living

The population of this group shows a higher proportion of people aged 65 and over than the parent Supergroup. Residents are more likely to live in communal establishments, detached properties and flats than the Supergroup, with a higher proportion of households living in socially rented accommodation.

5b1 – Delayed Retirement

The population of this group shows a lower proportion of people aged 90 and over than the parent Group, and households are more likely to live in flats, though are less likely to socially rent. There is a higher proportion of people who use public transport to get to work and they are more likely to work in the information and communication, and financial related industries.

5b2 – Communal Retirement

A distinguishing feature of this Subgroup is the high proportion of people living in communal establishments. The population of this Subgroup shows a higher proportion of people aged 90 and over than the parent Group. There is also a higher proportion of households living in terraced properties than the parent Group.

5b3 – Self-Sufficient Retirement

A lower proportion of people live in communal establishments than the parent Group. Compared with the Group a higher proportion of households live in terraced properties and households are more likely to live in socially rented accommodation.

C.1.6. Suburbanites

6 – Suburbanites

The population of this Supergroup is most likely to be located on the outskirts of urban areas. They are more likely to own their own home and to live in semi-detached or detached properties. The population tends to be a mixture of those above retirement age and middle-aged parents with school age children. The number of residents who are married or in civil-partnerships is above the national average. Individuals are likely to have higher-level qualifications than the national average, with the levels of unemployment in these areas being below the national average. All non-White ethnic groups have a lower representation when compared with the UK and the proportion of people born in the UK or Ireland is slightly higher. People are more likely to work in the information and communication, financial, public administration, and education sectors, and use private transport to get to work.

6a – Suburban Achievers

When compared with the parent Supergroup a higher proportion of households live in detached properties and flats, and are less likely to rent their accommodation or live in overcrowded conditions. People of Indian ethnicity are over-represented when compared with the Supergroup. Higher proportions of people have higher qualifications, and are more likely to work in the information and communication, and financial related industries.

6a1 – Indian Tech Achievers

All non-White ethnic groups are well represented in this Subgroup in comparison to the parent Group, people of Indian ethnicity being particularly well represented. There is a higher proportion of people born in other EU countries, and whose main language is not English or Welsh. Households are more likely to live in semi-detached properties. Compared with the parent Group there is a higher proportion of people working in the information and communication, and financial related industries, and workers using public transport.

6a2 – Comfortable Suburbia

The population of this group has a higher proportion of people aged 0 to 44 but a lower proportion aged 65 and over than the parent Group. Households are less likely to live in

semi-detached properties or flats, but more likely to live in detached or terraced properties.

6a3 – Detached Retirement Living

This Subgroup has a higher proportion of people aged 65 to 89 than the parent Group. There is a lower representation of all non-White ethnic groups in the Subgroup. Households are more likely to live in semi-detached properties.

6a4– Ageing in Suburbia

The population of this Subgroup has a higher proportion of people aged 65 and over than the parent Group. A much higher proportion of households live in flats, and households are more likely to live in privately rented accommodation. Households are also more likely to live in overcrowded conditions.

6b – Semi-Detached Suburbia

People in this group are slightly more likely to be divorced or separated than those in the Supergroup. Households are more likely to live in semi-detached and terraced properties, with a higher proportion of households renting their accommodation.

6b1 – Multi-Ethnic Suburbia

All the non-White ethnic groups are represented more highly in this Subgroup in comparison with the parent Group. There are also higher proportions of people born in the new EU countries and people whose main language is not English or Welsh. Households are more likely to live in semi-detached properties and to live in overcrowded conditions. A higher proportion of workers use public transport to commute to work.

6b2 – White Suburban Communities

Ethnic group representation, including persons with White ethnicity, is very similar to the parent Group. The population of this Subgroup has a lower proportion of people aged 65 and over than the parent Group. Households are more likely to live in detached or terraced properties, and to live in privately rented accommodation.

6b3 – Semi-Detached Ageing

This Subgroup has a higher proportion of people aged 65 to 89 than the parent Group. All non-White ethnic groups have a lower representation in this Subgroup when

compared with the parent Group. A higher proportion of households live in semi-detached properties, and own their own property.

6b4 – Older Workers and Retirement

There is a higher proportion of residents aged 65 to 89 within this Subgroup than the parent Group, and households are more likely to live in detached properties or flats. A higher proportion of households socially rent their accommodation, and a higher proportion are likely to live in overcrowded conditions.

C.1.7. Constrained City Dwellers

7 – Constrained City Dwellers

This Supergroup has a lower proportion of people aged 5 to 14 and a higher level aged 65 and over than nationally. It is more densely populated than the UK average. People are more likely to be single or divorced. There is a lower representation of all the non-White ethnic groups and of people who were born in other EU countries. There is a lower proportion of households with no children. Households are more likely to live in flats and to live in social rented accommodation, and there is a higher prevalence of overcrowding. There is a higher proportion of people whose day-to-day activities are limited, and lower qualification levels than nationally. There is a higher level of unemployment in the Supergroup. There are no particular industries in which workers are most likely to be employed, but some industries such as information and communication, and the education sector are underrepresented.

7a – Challenged Diversity

The population of this group have a higher level of people aged 0 to 14 in comparison with the Supergroup. All non-White ethnic groups have a higher representation than nationally, especially people who have mixed ethnicity. A higher proportion of households live in terraced properties, and are more likely to live in private rented accommodation when compared with the Supergroup. Car ownership is generally higher than the Supergroup, and people are more likely to be employed in information and communication related industries.

7a1 – Transitional Eastern European Neighbourhood

All non-White ethnic groups have a lower representation in this Subgroup when compared with the parent Group. A higher proportion of people were born in the new EU countries and there is a higher proportion whose main language is not English or Welsh. Households are more likely to live in detached properties and a greater proportion live in privately rented accommodation when compared with the parent Group. People are more likely to work in the agriculture and manufacturing related industries, and to use private transport to get to work.

7a2 – Hampered Aspiration

The population of this Subgroup has a lower representation of people of mixed ethnicity or of Black ethnicity when compared with the parent Group. A higher proportion of households live in terraced houses and in privately rented accommodation when compared with the parent Group. A higher proportion of people work in the information and communication, financial, and public administration related industries.

7a3 – Multi-Ethnic Hardship

The age make-up of this Subgroup is higher in the 5 to 14 age group when compared with the parent Group. Whilst there are higher proportions of people of mixed or Black ethnicity, all ethnic groups are well represented, though a lower proportion of people were born in other EU countries. Households were more likely to live in semi-detached properties and were more likely to live in socially rented accommodation. Workers were more likely to be employed in transport or storage industries.

7b – Constrained Flat Dwellers

This group is characterised by people living in flats, with a higher proportion living in socially rented accommodation than for the Supergroup. Ethnic groups generally have a similar representation as for the Supergroup, persons of mixed ethnicity are underrepresented. There is a lower proportion of households with two or more cars.

7b1 – Eastern European Communities

The population of this group has a higher proportion of people aged 0 to 14 than the parent Group and a lower proportion aged 65 and over. A higher proportion of people were born in the new EU countries. There is also a higher proportion of people whose main language is not English or Welsh. Households are more likely to live in socially

rented accommodation, with workers more likely to be employed in manufacturing industries.

7b2 – Deprived Neighbourhoods

The age structure for this Subgroup is very similar to the parent Group, with a lower proportion of people born in the new EU countries, and whose main language is not English or Welsh. Households are less likely to own their property and more likely to socially rent. There is a higher proportion of unemployed people.

7b3 – Endeavouring Flat Dwellers

In this Subgroup there is a lower proportion of people born in new EU countries. There is also a lower proportion of people whose main language is not English or Welsh. In comparison with the parent Group households are more likely to have no children, and to own their accommodation, but they are also more likely to be living in privately rented accommodation. People are less likely to be unemployed and there is a higher proportion of people working in the information and communication, finance and public administration, and education related sectors.

7c – White Communities

The population of this group are more likely to be white when compared with the parent Supergroup, with a lower representation of all other ethnic groups, and a lower proportion of people born in other EU countries. There is a higher proportion of households with non-dependent children, with households more likely to be living in semi-detached and terraced properties, and owning their own accommodation.

7c1 – Challenged Transitionaries

Households in this Subgroup are more likely to live in terraced properties than those in the parent Group. There is a lower proportion social renting and a higher proportion renting privately compared with the parent Group. People are less likely to be unemployed and are more likely to work in the information and communication, and financial related industries.

7c2 – Constrained Young Families

The population of this Subgroup shows a higher proportion of people aged 0 to 14 than the parent Group and a lower proportion of families with no children. Households are more likely to live in terraced properties and to socially rent their accommodation. There

is a higher proportion of people who are unemployed and those in employment are more likely to work in the accommodation or food service activities industries.

7c3 – Outer City Hardship

The population of this Subgroup has a higher proportion of people aged 65 and over when compared with the parent Group. Households are more likely to live in detached and semi-detached properties, and private renting is slightly more prevalent than the parent Group. There is a higher proportion of households with two or more cars and a lower proportion of people who use public transport to get to work.

7d – Ageing City Dwellers

The population of this group shows a higher proportion of people aged 65 and over when compared with the parent Supergroup, and residents are more likely to live in communal establishments and less likely to be single. There is a higher proportion of households living in detached properties and flats. A lower proportion of people are unemployed.

7d1 – Ageing Communities and Families

The age profile of this Subgroup shows a higher proportion of people aged 0 to 14 and a lower proportion aged 90 and over. People are less likely to live in communal establishments than the parent Group. There is a higher proportion of households living in detached and semi-detached properties, and households owning their own property. Households are more likely to have two or more cars. There is a higher proportion of people working in the information and communication, and education related sectors.

7d2 – Retired Independent City Dwellers

When compared with the parent Group there is a higher proportion of households living in flats, and in socially rented accommodation. There is a lower proportion of households with two or more cars, but higher unemployment amongst residents.

7d3 – Retired Communal City Dwellers

The population of this Subgroup shows a lower proportion of people aged 0 to 14 than the parent Group but a higher proportion aged 90 and over. A much higher proportion of people live in communal establishments. When compared with the parent Group a higher proportion of households live in detached, semi-detached or terraced properties, privately rent their accommodation, and have two or more cars.

7d4 – Retired City Hardship

The age profile of this Subgroup shows a higher proportion of the population aged 65 and over, though residents are less likely to live in communal establishments or to be single. There is a lower proportion of households with no children and non-dependent children. Households are more likely to live in flats, and more likely to rent socially. People in this Subgroup are less likely to have qualifications, and due to the age structure are less likely to work in any of the identified industry categories.

C.1.8. Hard-Pressed Living

8 – Hard-Pressed Living

The population of this group is most likely to be found in urban surroundings, predominately in northern England and southern Wales. There is less non-White ethnic group representation than elsewhere in the UK, and a higher than average proportion of residents born in the UK and Ireland. Rates of divorce and separation are above the national average. Households are more likely to have non-dependent children and are more likely to live in semi-detached or terraced properties, and to socially rent. There is a smaller proportion of people with higher level qualifications, with rates of unemployment above the national average. Those in employment are more likely to be employed in the mining, manufacturing, energy, wholesale and retail, and transport related industries.

8a – Industrious Communities

Age structure and ethnic group representation broadly reflects the parent Supergroup. There is a higher proportion of households living in detached and semi-detached properties, with slightly higher property ownership than for the Supergroup. Industrious communities have a broadly similar demographic to the Supergroup in terms of age group, occupation and population density, however slightly less overcrowding exists in this group. Ownership of two or more cars or vans is also marginally higher.

8a1 – Industrious Transitions

The Subgroup is broadly similar to the parent Group in terms of age groups and ethnic diversity. Compared with the parent Group, social renting is less common while a higher proportion of residents live in detached properties. This Subgroup exhibits slightly

higher proportions of residents in the information and communication, and financial related industries, together with slightly higher levels of educational qualifications.

8a2 – Industrious Hardship

Age structure and ethnicity for this Subgroup are consistent with the parent Group.
This Subgroup has a higher proportion of residents who live in semi-detached properties, and social renting is more common, with marginally higher overcrowding. There is a smaller proportion of people in the information and communication, and financial related industries, and slightly higher levels of unemployment.

8b – Challenged Terraced Workers

A key difference with this group compared with the parent Supergroup is the dominance of terraced housing over other types. Ownership of two or more cars and non-White ethnic group representation is also lower. The group has a similar age structure to the Supergroup and similar employment characteristics.

8b1 – Deprived Blue-Collar Terraces

Whilst for this Subgroup, the age structure is broadly similar to the parent Group, it is characterised by a higher degree of non-White ethnic group representation, and higher levels of households in private rented accommodation. There are also marginally higher levels of educational qualifications with employment in the information and communication, financial, and education sectors more prevalent than with the parent Group.

8b2 – Hard-Pressed Rented Terraces

For this Subgroup, age structure and representation of ethnic groups are broadly similar to the parent Group. There is a higher proportion of households living in semi-detached properties, and a noticeably higher proportion living in socially rented accommodation. There is also a slightly higher use of public transport for commuting to work.

8c – Hard-Pressed Ageing Workers

Residents who live in this group have a broadly similar age structure to the Supergroup, though a smaller proportion of young people and higher proportion of older people. There is less non-While ethnic group representation than with the parent Supergroup. Employment characteristics for this group closely reflect those for the Supergroup.

8c1 – Ageing Industrious Workers

This Subgroup is characterised by a slightly older age profile, and a higher proportion of mixed ethnicity of the residents compared with the parent Group. Households are more likely to live in detached properties, and to live in privately rented accommodation. Residents have higher educational qualifications, whilst workers are more likely to be employed in the information and communication, and financial related industries.

8c2 – Ageing Rural Industry Workers

The age structure is very similar to the parent Group, though there is lower non-White ethnic group representation. Households are slightly more likely to live detached or terraced housing, and to be living in private rented accommodation when compared with the parent Group. This Subgroup contains considerably higher proportions of workers in agriculture, forestry and fishing occupations than the parent Group, with a higher proportion of people walking and/or cycling to work.

8c3 – Renting Hard-Pressed Workers

This Subgroup has smaller proportions of older residents, and ethnic group representation is reasonably similar to the parent Group. A higher proportion of households live in semi-detached properties and flats compared with the parent Group, and households are more likely to live in socially rented accommodation. Household overcrowding and unemployment is marginally higher than the parent Group. Workers are more likely to work in transport or storage industries, and to use public transport to travel to work.

8d – Migration and Churn

This group has a higher proportion of children aged 0 to 14 than the Supergroup, with a higher representation of non-While ethnic groups. Households are more likely to live in terraced houses or flats, and to socially rent their property. Unemployment is noticeably higher than for the Supergroup, and people are more likely to be employed in the tertiary industry (service) sector, and use public transport to get to work.

8d1 – Young Hard-Pressed Families

The age structure for this Subgroup largely reflects the parent Group. There is a lower non-White ethnic group representation, particularly for people of mixed ethnicity. Households in socially rented accommodation are more prevalent than with the parent

Group, and unemployment is higher. Certain industries such as the information and communication related industries are underrepresented with this Subgroup.

8d2 – Hard-Pressed Ethnic Mix

This Subgroup is generally characterised by higher levels of non-White ethnic group representation, particularly of persons of mixed and Black ethnicity, together with an older age structure than for the parent Group. Households are more likely to live in flats, and more likely to own their own property. There are also a higher proportion of workers in the information and communication, and financial related industries, whilst unemployment is marginally lower than for the parent Group.

8d3 – Hard-Pressed European Settlers

The key characteristic of this Subgroup is the higher proportion of residents who were born in new EU countries compared with the parent Group, and a younger age structure. Main language not English or Welsh is also more prevalent. Households in this Subgroup are more likely to live in privately rented accommodation when compared with the parent Group. Unemployment is marginally lower than for the parent Group, with those in employment more likely to travel to work by walking or cycling, and employed in agriculture and manufacturing related industries.

C.2. Towns and cities used for geographic distribution analysis

The following towns and cities were used to test how the distributions of the 167 variables initially selected to be part of the 2011 OAC, and the final selection of 60 varied between different urban areas in the UK. The results of this analysis are discussed in Sections 7.2.2.4 and 7.2.3.

Aberdeen
Belfast
Birmingham
Bradford
Brighton and Hove
Bristol
Cardiff
Coventry
Dundee
Edinburgh
Glasgow
Kingston upon Hull
Leeds
Leicester
Liverpool
London
Manchester
Newcastle upon Tyne
Nottingham
Plymouth
Sheffield
Southampton
Stoke-on-Trent
Swansea
Wolverhampton

C.3. Distribution plots of the 2011 OAC's 167 initially selected variables

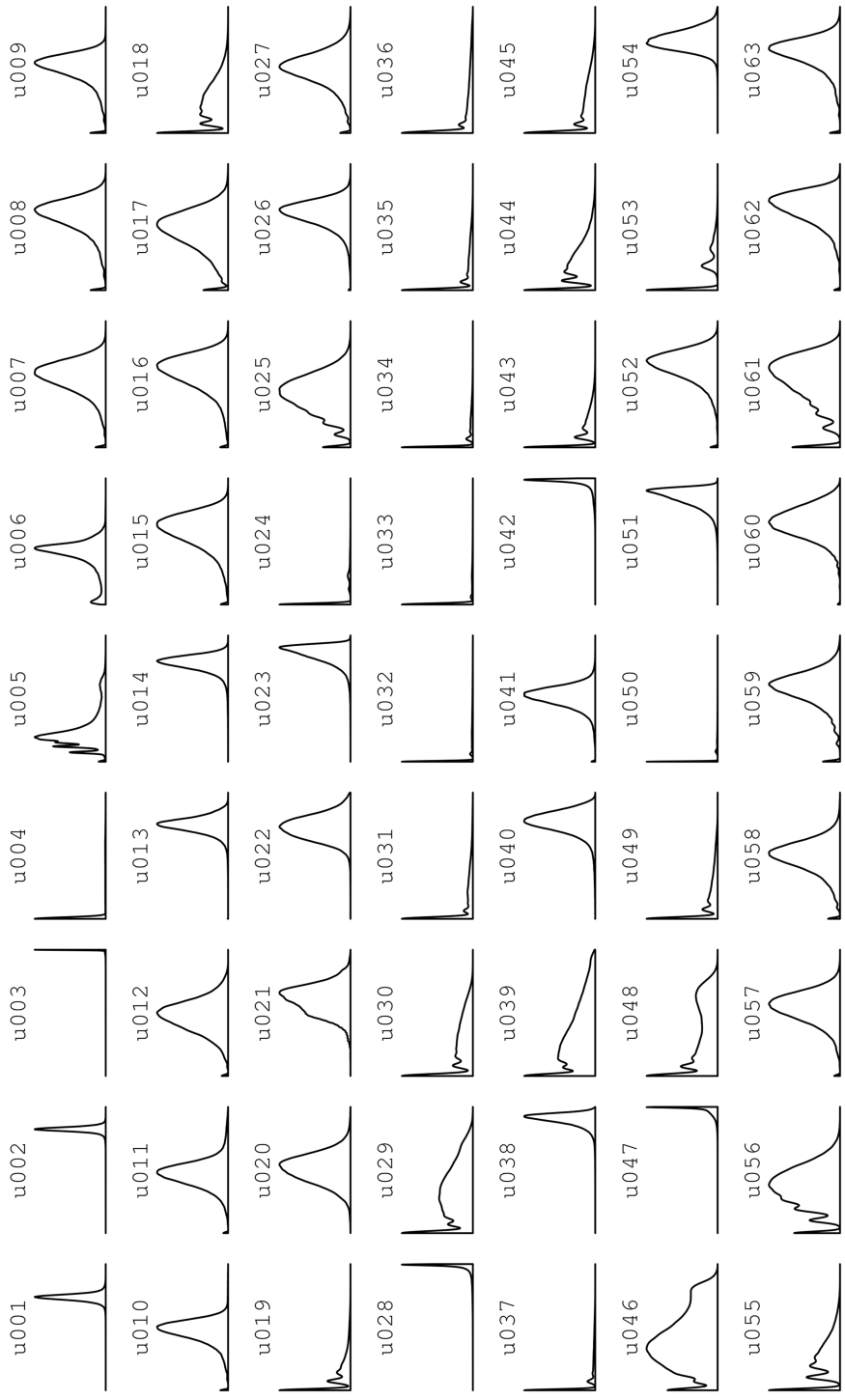


Figure C.1 (Part 1): Distribution plots of the 2011 OAC's 167 initially selected variables

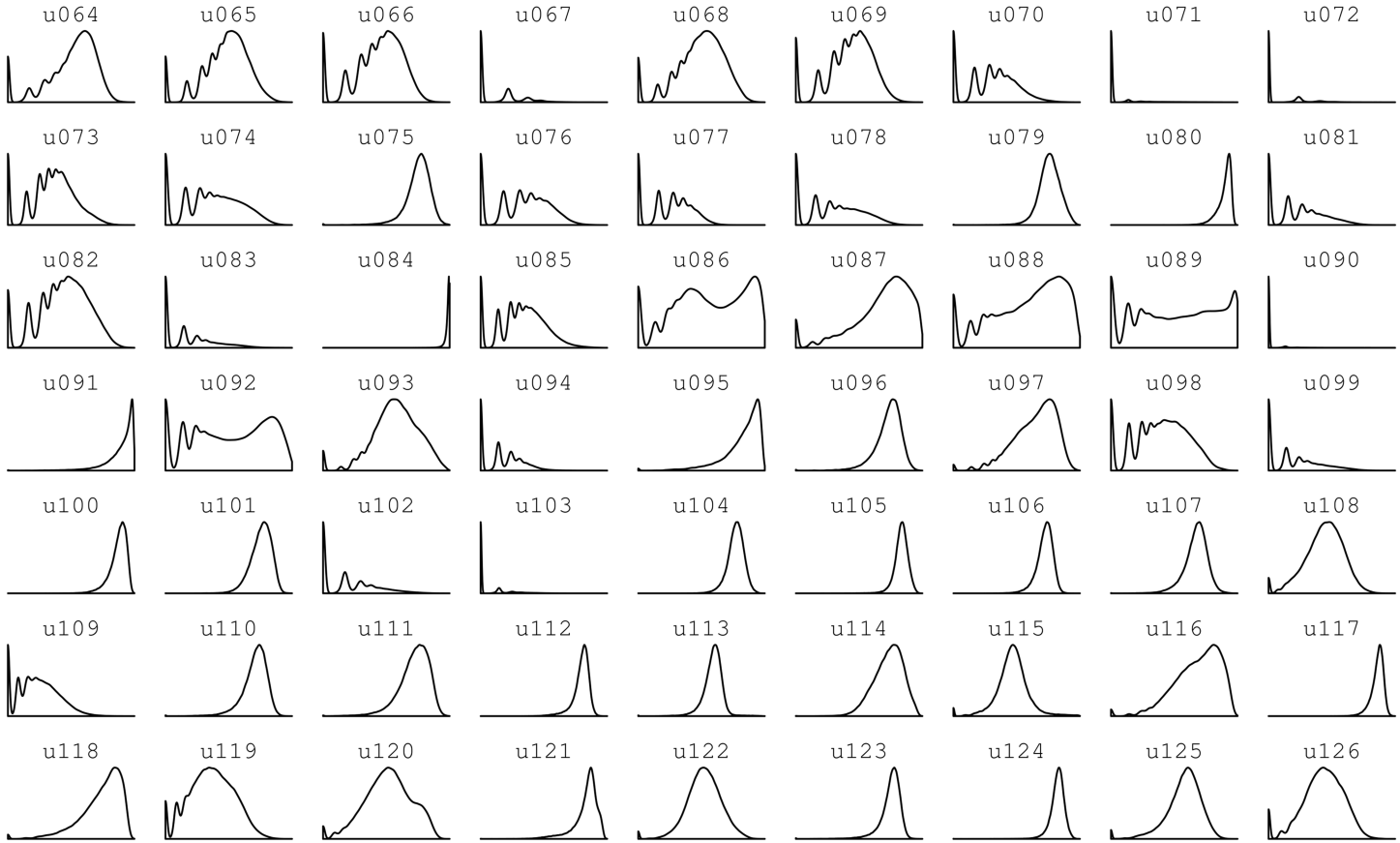


Figure C.1 (Part 2): Distribution plots of the 2011 OAC's 167 initially selected variables

Figure C.1 (Part 3): Distribution plots of the 2011 OAC's 167 initially selected variables

C.4. Histograms of the potential 2011 OAC datasets distributions

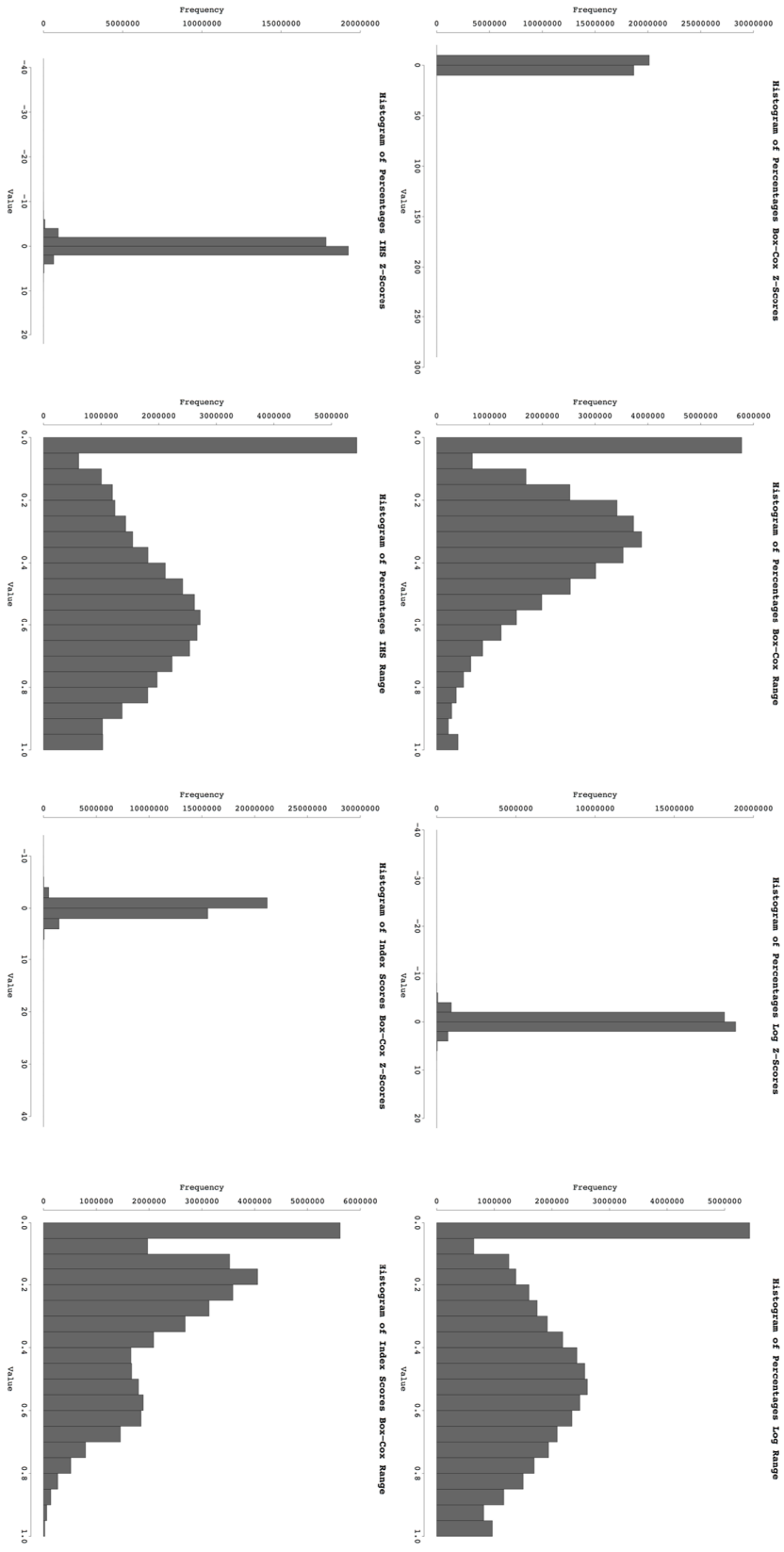


Figure C.2 (Part 1): Histograms of the potential 2011 OAC datasets distributions

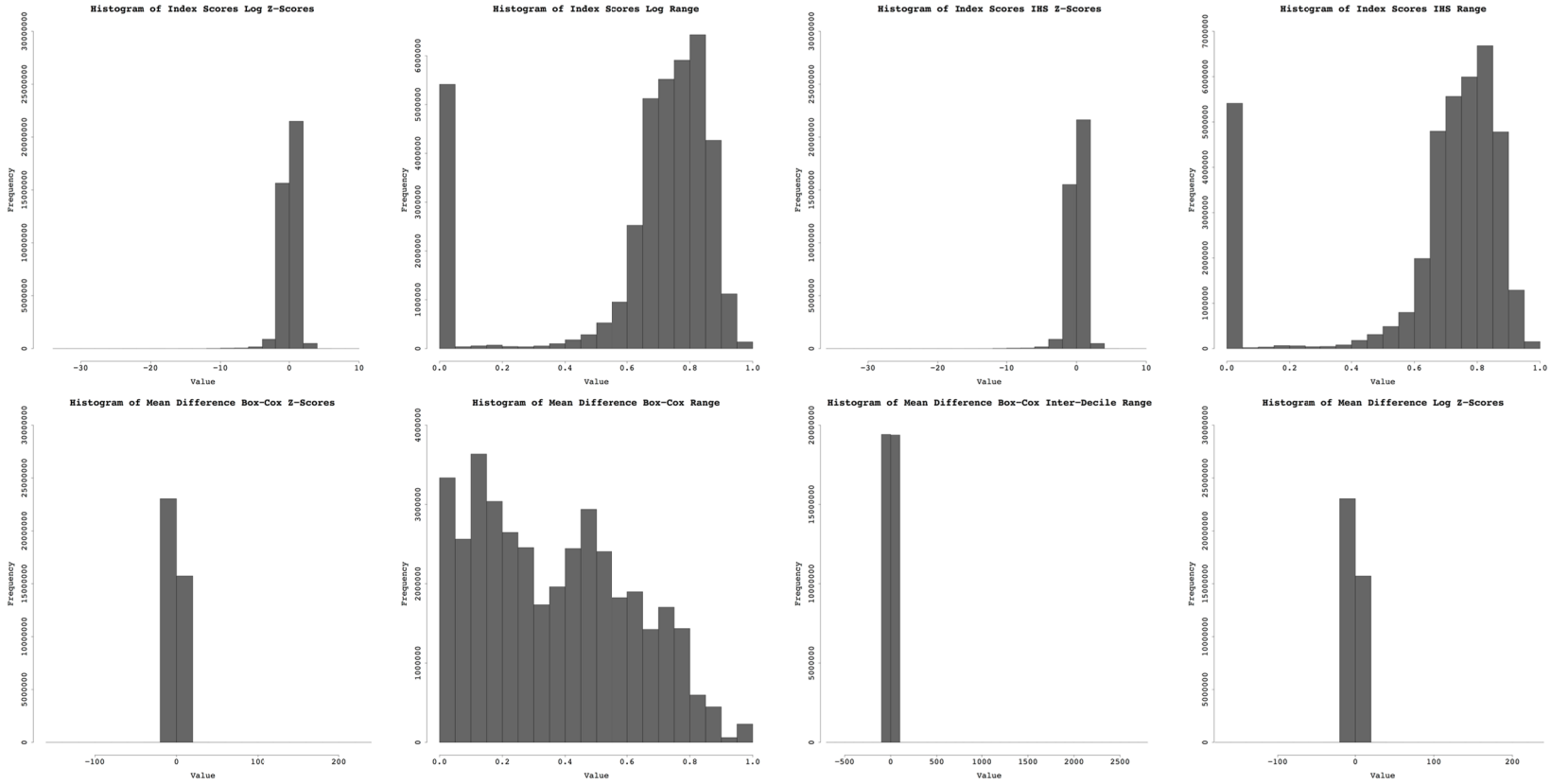


Figure C.2 (Part 2): Histograms of the potential 2011 OAC datasets distributions

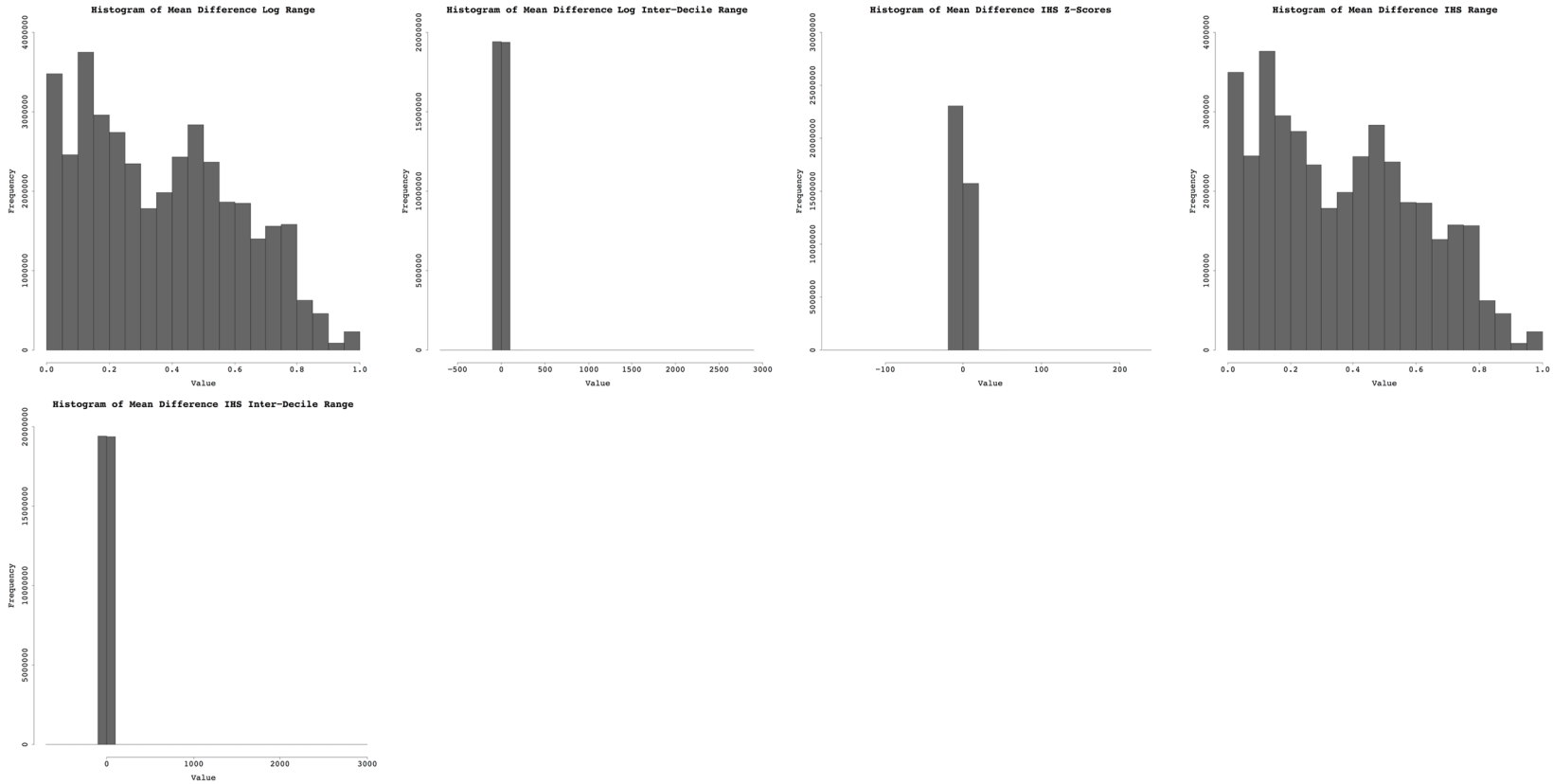


Figure C.2 (Part 3): Histograms of the potential 2011 OAC datasets distributions

C.5. 2011 OAC final variable selection rationale

Table C.1: Rationale for the 2011 OAC final variable selection

Code	Variable Name	Kept, rejected or merged
u001	Males	Rejected due to negative correlation with u002 and no variation across the UK
u002	Females	Rejected due to negative correlation with u001 and no variation across the UK
u003	Persons living in a household	Rejected due to negative correlation with u004 and highly skewed distribution across the UK
u004	Persons living in a communal establishment	Kept, despite highly skewed distribution across the UK and WCSS analysis suggesting it performs badly in a clustering solution, to identify areas where high concentrations of communal establishments are located
u005	Area size (in hectares)	Rejected as u006 provided a better indication of population density across different areas of the UK
u006	Number of persons per hectare	Kept as it provided a good indication of the variation in population density across different areas in the UK
u007	Persons aged 0 to 4	Kept as it provided an indicator of the pre-school age population in the UK
u008	Persons aged 5 to 9	Merged with u009 to create a single indicator of the school age population in the UK
u009	Persons aged 10 to 14	Merged with u008 to create a single indicator of the school age population in the UK
u010	Persons aged 15 to 19	Rejected due to significant positive correlation with u115
u011	Persons aged 20 to 24	Rejected due to significant positive correlation with u022
u012	Persons aged 25 to 29	Merged with u013 to make the resulting composite variable less correlated with other variables

Code	Variable Name	Kept, rejected or merged
u013	Persons aged 30 to 44	Merged with u012 to make the resulting composite variable less correlated with other variables
u014	Persons aged 45 to 59	Merged with u015 to make the resulting composite variable less correlated with other variables
u015	Persons aged 60 to 64	Merged with u014 to make the resulting composite variable less correlated with other variables
u016	Persons aged 65 to 74	Merged with u017 and u018 to make the resulting composite variable less correlated with other variables
u017	Persons aged 75 to 84	Merged with u016 and u018 to make the resulting composite variable less correlated with other variables
u018	Persons aged 85 to 89	Merged with u016 and u017 to make the resulting composite variable less correlated with other variables
u019	Persons aged 90 and over	Kept as it provided a good indicator of the older population in the UK
u020	Mean age	Rejected due to significant correlation with numerous other variables and had limited differentiating power across the UK
u021	Median age	Rejected due to significant correlation with numerous other variables and had limited differentiating power across the UK
u022	Persons aged over 16 who are single	Kept as it provided a good indicator for the younger population in the UK
u023	Persons aged over 16 who are married	Merged with u024 to make the resulting composite variable less skewed and correlated with other variables
u024	Persons aged over 16 who are in a registered same-sex civil partnership	Merged with u023 to make the resulting composite variable less skewed and correlated with other variables
u025	Persons aged over 16 who are separated	Merged with u026 to create composite variable that represented all separated and divorced individuals in the UK
u026	Persons aged over 16 who are divorced or formerly in a same-sex civil partnership which is now legally dissolved	Merged with u025 to create composite variable that represented all separated and divorced individuals in the UK

Code	Variable Name	Kept, rejected or merged
u027	Persons aged over 16 who are widowed or a surviving partner from a same-sex civil partnership	Rejected due to significant correlation with numerous other variables and had limited differentiating power across the UK
u028	Persons who are white British and Irish	Merged with u029 to make the resulting composite variable less skewed
u029	Persons who are other white	Merged with u028 to make the resulting composite variable less skewed and correlated with other variables
u030	Persons who have mixed ethnicity or are from multiple ethnic groups	Kept as it provided good differentiation across different parts of the UK and offered insight into the ethnic identity of urban areas
u031	Persons who are Asian/Asian British: Indian	Kept as it provided good differentiation across different parts of the UK and offered insight into the ethnic identity of urban areas
u032	Persons who are Asian/Asian British: Pakistani	Kept as it provided good differentiation across different parts of the UK and offered insight into the ethnic identity of urban areas
u033	Persons who are Asian/Asian British: Bangladeshi	Kept as it provided good differentiation across different parts of the UK and offered insight into the ethnic identity of urban areas despite its highly skewed distribution
u034	Persons who are Asian/Asian British: Chinese	Merged with u035 to make the resulting composite variable less skewed and represent a greater share of the population
u035	Persons who are Asian/Asian British: Other	Merged with u034 to make the resulting composite variable less skewed and represent a greater share of the population
u036	Persons who are Black/African/Caribbean/ Black British	Kept as it provided good differentiation across different parts of the UK and offered insight into the ethnic identity of urban areas
u037	Persons who are Arab or are from another ethnic group	Kept as it offered insight into the ethnic identity of urban areas in the UK
u038	Persons who are Christian	Rejected due to significant correlation with multiple other variables
u039	Persons who are from another religion	Rejected due to significant correlation with numerous other variables and because of a skewed distribution
u040	Persons who have no religion	Rejected as no other religion variables (u038 and u039) were retained

Code	Variable Name	Kept, rejected or merged
u041	Persons who did not state their religion	Rejected as no other religion variables (u038 and u039) were retained
u042	Persons whose country of birth is the United Kingdom	Merged with u043 to make the resulting composite variable less skewed and reduce significant correlation with other variables
u043	Persons whose country of birth is Ireland	Merged with u042 to make the resulting composite variable less skewed and reduce significant correlation with other variables
u044	Persons whose country of birth is in the old EU (pre 2004 accession countries)	Kept as it provided an indicator of the population of the UK from pre-accession EU countries
u045	Persons whose country of birth is in the new EU (post 2004 accession countries)	Kept as it provided an indicator of the population of the UK from post-accession EU countries
u046	Persons whose country of birth is not the UK, Ireland or EU countries	Rejected due to significant correlation with numerous other variables, including a strong negative correlation with u042 and u028
u047	Persons whose main language is English or their main language is not English but can speak English very well	Rejected due to significant correlation with numerous other variables and as the composite variable created from u049 and u050 provides an indicator of language skills
u048	Persons whose main language is not English but can speak English well	Rejected due to significant correlation with numerous other variables and as the composite variable created from u049 and u050 provides an indicator of language skills
u049	Persons whose main language is not English and cannot speak English well	Merged with u050 due to high correlation and to make the resulting composite variable an indicator of persons without good English language skills
u050	Persons whose main language is not English and cannot speak English	Merged with u049 due to high correlation and to make the resulting composite variable an indicator of persons without good English language skills
u051	Households that only contain Persons aged over 16 who are living in a couple: Married	Rejected due to significant correlation with numerous other variables, including a strong positive correlation with u023
u052	Households that only contain Persons aged over 16 who are living in a couple: Cohabiting (opposite-sex)	Rejected due to significant positive correlation with u065

Code	Variable Name	Kept, rejected or merged
u053	Households that only contain Persons aged over 16 who are living in a couple: In a registered same-sex civil partnership or cohabiting (same-sex)	Rejected due to significantly skewed distribution and as no other 'living in couple' variables (u051 and u052) were retained
u054	Households that only contain Persons aged over 16 who are not living in a couple: Single (never married or never registered a same-sex civil partnership)	Rejected due to significant correlation with numerous other variables, including a strong positive correlation with u022
u055	Households that only contain Persons aged over 16 who are not living in a couple: Married or in a registered same-sex civil partnership	Rejected due to significant negative correlation with u028 and u042
u056	Households that only contain Persons aged over 16 who are not living in a couple: Separated (but still legally married or still legally in a same-sex civil partnership)	Rejected due to significant positive correlation with u025
u057	Households that only contain Persons aged over 16 who are not living in a couple: Divorced or formerly in a same-sex civil partnership which is now legally dissolved	Rejected due to significant positive correlation with u026
u058	Households that only contain Persons aged over 16 who are not living in a couple: Widowed or surviving partner from a same-sex civil partnership	Rejected due to significant correlation with numerous other variables
u059	One person households: Aged 65 and over	Rejected due to significant correlation with numerous other variables
u060	One person households: Other	Rejected due to significant correlation with numerous other variables

Code	Variable Name	Kept, rejected or merged
u061	One family households: All aged 65 and over	Rejected due to significant correlation with numerous other variables
u062	One family households: Married or same-sex civil partnership couple with no children	Merged with u065 to create a composite variable that had better differentiation across the UK and represented a greater share of the population
u063	One family households: Married or same-sex civil partnership couple with dependent children	Rejected due to significant correlation with numerous other variables
u064	One family households: Married or same-sex civil partnership couple with non-dependent children	Merged with u067 and u069 to create a composite variable that had better differentiation across the UK and represented a greater share of the population
u065	One family households: Cohabiting couple with no children	Merged with u062 to create a composite variable that had better differentiation across the UK and represented a greater share of the population
u066	One family households: Cohabiting couple with dependant children	Rejected due to limited variation across the UK
u067	One family households: Cohabiting couple with non-dependant children	Merged with u064 and u069 to create a composite variable that had better differentiation across the UK and represented a greater share of the population
u068	One family households: Lone parent with dependant children	Rejected due to significant correlation with numerous other variables
u069	One family households: Lone parent with non-dependant children	Merged with u064 and u067 to create a composite variable that had better differentiation across the UK and represented a greater share of the population
u070	Other household types: With dependent children	Rejected as no other 'dependent children' variables (u063, u066, u068 and u074) were retained
u071	Other household types: All full-time students	Kept as it provided an indicator for student only households which represents a large and important section of society
u072	Other household types: All aged 65 and over	Rejected due to due to highly skewed distribution, limited variation across the UK and as WCSS analysis suggests it performs badly in a clustering solution
u073	Other household types: Other	Rejected due to limited descriptive power

Code	Variable Name	Kept, rejected or merged
u074	Households with no adults in employment: With dependent children	Rejected due to significant correlation with numerous other variables
u075	Households with no adults in employment: No dependent children	Rejected due to significant correlation with numerous other variables
u076	Households with lone parent in part-time employment	Rejected due to limited variation across the UK
u077	Households with lone parent in full-time employment	Rejected due to limited variation across the UK
u078	Households with lone parent not in employment	Rejected due to significant correlation with numerous other variables
u079	One person ethnic household	Rejected due to significant correlation with numerous other variables
u080	Household members all have the same ethnic group	Rejected due to significant correlation with numerous other variables
u081	Households with different ethnic groups between the generations only	Rejected due to significant correlation with numerous other variables
u082	Households with different ethnic groups within partnerships (whether or not different ethnic groups between generations)	Rejected due to limited descriptive power
u083	Households with any other combination of multiple ethnic groups	Rejected due to significant correlation with numerous other variables
u084	Household spaces with at least one usual resident	Rejected due to significant negative correlation with u085 and highly skewed distribution
u085	Household spaces with no usual residents	Rejected due to significant negative correlation with u084 and due to limited variation across the urban areas of the UK
u086	Households who live in a detached house or bungalow	Kept as the variation in the types of housing directly influence the physical characteristics of an area

Code	Variable Name	Kept, rejected or merged
u087	Households who live in a semi-detached house or bungalow	Kept as the variation in the types of housing directly influence the physical characteristics of an area
u088	Households who live in a terrace or end-terrace house	Kept as the variation in the types of housing directly influence the physical characteristics of an area
u089	Households who live in a flat	Kept as the variation in the types of housing directly influence the physical characteristics of an area
u090	Households who live in a caravan or other mobile or temporary structure	Rejected due to limited variation across the UK and as WCSS analysis suggests it performs badly in a clustering solution
u091	Households who own or have shared ownership of property	Kept as it provided a good indicator of areas where the population can afford to own their own home and has high variation across the UK
u092	Households who are social renting	Kept as it provided a good indicator of areas where the population cannot afford to rent privately or own their own home and has high variation across the UK
u093	Households who are private renting	Kept as it provided a good indicator of areas where the population can afford to rent privately and has high variation across the UK
u094	Households who are living rent free	Rejected due to due limited variation across the UK and as WCSS analysis suggests it performs badly in a clustering solution
u095	Households who have two or more rooms than required	Rejected due to significant correlation with numerous other variables and as the composite variable created from u098 and u099 provides an indicator of household crowding
u096	Households who have one more room than required	Rejected as the composite variable created from u098 and u099 provides an indicator of household crowding
u097	Households who have the required number of rooms	Rejected due to significant correlation with numerous other variables and as the composite variable created from u098 and u099 provides an indicator of household crowding
u098	Households who have one fewer room than required	Merged with u099 to make the resulting composite variable less correlated with other variables and provide an indicator of overcrowded households
u099	Households who have two fewer or less rooms than required	Merged with u098 to make the resulting composite variable less correlated with other variables and provide an indicator of overcrowded households
u100	Households with up to 0.5 persons per room	Rejected due to significant correlation with numerous other variables and as the composite variable created from u098 and u099 provides an indicator of household crowding

Code	Variable Name	Kept, rejected or merged
u101	Households with over 0.5 and up to 1.0 persons per room	Rejected due to significant correlation with numerous other variables and as the composite variable created from u098 and u099 provides an indicator of household crowding
u102	Households with over 1.0 and up to 1.5 persons per room	Rejected due to significant correlation with numerous other variables and as the composite variable created from u098 and u099 provides an indicator of household crowding
u103	Households with over 1.5 persons per room	Rejected due to significant correlation with numerous other variables and as the composite variable created from u098 and u099 provides an indicator of household crowding
u104	Day-to-day activities limited a lot or a little Standardised Illness Ratio	Kept as it provided a good indicator of population health in the UK
u105	Persons in very good health	Rejected due to significant correlation with numerous other variables and as u104 provides information on the population health in the UK
u106	Persons in good health	Rejected as u104 provides information on the population health in the UK
u107	Persons in fair health	Rejected due to significant correlation with numerous other variables and as u104 provides information on the population health in the UK
u108	Persons in bad health	Rejected due to significant correlation with numerous other variables and as u104 provides information on the population health in the UK
u109	Persons in very bad health	Rejected as WCSS analysis suggests it performs badly in clustering solutions and as u104 provides information on the population health in the UK
u110	Persons providing unpaid care	Kept as it provided a non-direct measure of population health in the UK
u111	Persons aged over 16 who have no qualifications	Rejected due to significant correlation with numerous other variables, including a strong negative correlation with u114
u112	Persons aged over 16 whose highest level of qualification is Level 1, Level 2 or Apprenticeship	Kept as it provided an indication of less well-off areas in the UK where the population had a basic education
u113	Persons aged over 16 whose highest level of qualification is Level 3 qualifications	Kept as it provided an indication of reasonably well-off areas in the UK where the population had an advanced education

Code	Variable Name	Kept, rejected or merged
u114	Persons aged over 16 whose highest level of qualification is Level 4 qualifications and above	Kept as it provided a good indication of better off areas in the UK with people who had a very good education
u115	Persons aged over 16 who are schoolchildren or full-time students	Kept as it provided a good indicator for areas that contain students no matter what their living arrangements, unlike u071
u116	Households with no cars or vans	Rejected due to significant correlation with numerous other variables, including a strong negative correlation with u118
u117	Households with 1 car or van	Rejected as limited variation across the UK and as u118 provides information on the vehicles in households
u118	Households with 2 or more cars or vans	Kept as it provided an indicator for potentially better-off households
u119	Persons aged between 16 and 74 who work mainly at or from home	Rejected as limited variation across the UK and as u120 to u122 provides information on individuals method of travel to work
u120	Persons aged between 16 and 74 who use public transport to get to work	Kept as it provided a good indication of areas that contain commuters
u121	Persons aged between 16 and 74 who use private transport to get to work	Kept as it provided a good indication of areas that contain commuters
u122	Persons aged between 16 and 74 who walk, cycle or use an alternative method to get to work	Kept as it provided a good indication of areas that contain workers that live close to their place of work
u123	Persons aged between 16 and 74 who are economically active: Part-time employees	Rejected as the composite variable created from u137 and u138 provides an indicator of part-time employment
u124	Persons aged between 16 and 74 who are economically active: Full-time employees	Rejected as the composite variable created from u139 and u140 provides an indicator of full-time employment
u125	Persons aged between 16 and 74 who are economically active: Self-employed	Rejected due to limited descriptive power
u126	Persons aged between 16 and 74 who are economically active: Unemployed	Kept as it identified individuals unemployed in an area but who were seeking employment in the week prior to the 27th March 2011 (Census day)

Code	Variable Name	Kept, rejected or merged
u127	Persons aged between 16 and 74 who are economically active: Full-time student	Rejected due to significant positive correlation with u115
u128	Persons aged between 16 and 74 who are economically inactive: Retired	Rejected due to significant correlation with numerous other variables, including a strong positive correlation with u016
u129	Persons aged between 16 and 74 who are economically inactive: Student (including full-time students)	Rejected due to significant positive correlation with u115
u130	Persons aged between 16 and 74 who are economically inactive: Looking after home or family	Rejected due to limited variation across the UK
u131	Persons aged between 16 and 74 who are economically inactive: Long-term sick or disabled	Rejected due to significant correlation with numerous other variables and as u104 provides information on the population health in the UK
u132	Persons aged between 16 and 74 who are economically inactive: Other	Rejected as limited variation across the UK
u133	Persons aged between 16 and 24 who are unemployed	Rejected as u126 provides information on the unemployment levels in the UK
u134	Persons aged between 50 and 74 who are unemployed	Rejected as u126 provides information on the unemployment levels in the UK
u135	Persons aged between 16 and 74 who have never worked	Rejected as u126 provides information on the unemployment levels in the UK
u136	Persons aged between 16 and 74 who are long-term unemployed	Rejected as u126 provides information on the unemployment levels in the UK
u137	Employed persons aged between 16 and 74: Part-time working 15 hours or less	Merged with u138 to create a composite variable that identified all part-time workers in an area
u138	Employed persons aged between 16 and 74: Part-time working 16 to 30 hours	Merged with u137 to create a composite variable that identified all part-time workers in an area
u139	Employed persons aged between 16 and 74: Full-time working 31 to 48 hours	Merged with u140 to create a composite variable that identified all full-time workers in an area

Code	Variable Name	Kept, rejected or merged
u140	Employed persons aged between 16 and 74: Full-time working 49 or more hours	Merged with u139 to create a composite variable that identified all full-time workers in an area
u141	Employed persons aged between 16 and 74 industry: Agriculture, forestry and fishing	Kept as it provided a good indicator of different types if industry individuals in the UK work in
u142	Employed persons aged between 16 and 74 industry: Mining and quarrying	Merged with u146 to create a composite variable that provided a good indicator of different types if industry individuals in the UK work in
u143	Employed persons aged between 16 and 74 industry: Manufacturing	Kept as it provided a good indicator of different types if industry individuals in the UK work in
u144	Employed persons aged between 16 and 74 industry: Electricity, gas, steam and air conditioning supply	Merged with u145 to create a composite variable that provided a good indicator of different types if industry individuals in the UK work in
u145	Employed persons aged between 16 and 74 industry: Water supply; sewerage, waste management and remediation activities	Merged with u144 to create a composite variable that provided a good indicator of different types if industry individuals in the UK work in
u146	Employed persons aged between 16 and 74 industry: Construction	Merged with u142 to create a composite variable that provided a good indicator of different types if industry individuals in the UK work in
u147	Employed persons aged between 16 and 74 industry: Wholesale and retail trade; repair of motor vehicles and motor cycles	Kept as it provided a good indicator of different types if industry individuals in the UK work in
u148	Employed persons aged between 16 and 74 industry: Transport and storage	Kept as it provided a good indicator of different types if industry individuals in the UK work in
u149	Employed persons aged between 16 and 74 industry: Accommodation and food service activities	Kept as it provided a good indicator of different types if industry individuals in the UK work in
u150	Employed persons aged between 16 and 74 industry: Information and communication	Merged with u153 to create a composite variable that provided a good indicator of different types if industry individuals in the UK work in

Code	Variable Name	Kept, rejected or merged
u151	Employed persons aged between 16 and 74 industry: Financial and insurance activities	Merged with u152 to create a composite variable that provided a good indicator of different types if industry individuals in the UK work in
u152	Employed persons aged between 16 and 74 industry: Real estate activities	Merged with u151 to create a composite variable that provided a good indicator of different types if industry individuals in the UK work in
u153	Employed persons aged between 16 and 74 industry: Professional, scientific and technical activities	Merged with u150 to create a composite variable that provided a good indicator of different types if industry individuals in the UK work in
u154	Employed persons aged between 16 and 74 industry: Administrative and support service activities	Kept as it provided a good indicator of different types if industry individuals in the UK work in
u155	Employed persons aged between 16 and 74 industry: Public administration and defence; compulsory social security	Kept as it provided a good indicator of different types if industry individuals in the UK work in
u156	Employed persons aged between 16 and 74 industry: Education	Kept as it provided a good indicator of different types if industry individuals in the UK work in
u157	Employed persons aged between 16 and 74 industry: Human health and social work activities	Kept as it provided a good indicator of different types if industry individuals in the UK work in
u158	Employed persons aged between 16 and 74 industry: Other industry	Rejected due to due limited variation across the UK and as WCSS analysis suggests it performs badly in a clustering solution
u159	Employed persons aged between 16 and 74 occupation: Managers, directors and senior officials	Rejected due to highly skewed distribution and limited variation across the UK
u160	Employed persons aged between 16 and 74 occupation: Professional occupations	Rejected due to significant correlation with numerous other variables, including a strong positive correlation with u114
u161	Employed persons aged between 16 and 74 occupation: Associate professional and technical occupations	Rejected due to highly skewed distribution and limited variation across the UK

Code	Variable Name	Kept, rejected or merged
u162	Employed persons aged between 16 and 74 occupation: Administrative and secretarial occupations	Rejected due to limited variation across the UK
u163	Employed persons aged between 16 and 74 occupation: Skilled trades occupations	Rejected due to limited variation across the UK
u164	Employed persons aged between 16 and 74 occupation: Caring, leisure and other service occupations	Rejected due to limited variation across the UK
u165	Employed persons aged between 16 and 74 occupation: Sales and customer service occupations	Rejected due to significant positive correlation with u147
u166	Employed persons aged between 16 and 74 occupation: Process, plant and machine operatives	Rejected due to significant negative correlation with u114 and because industry indicators rather than occupation were retained
u167	Employed persons aged between 16 and 74 occupation: Elementary occupations	Rejected due to significant negative correlation with u114

C.6. The Preliminary 2011 England and Wales OAC

Table C.2: The names for the Preliminary 2011 England and Wales OAC Supergroups, Groups and Subgroups

Supergroup	Group	Subgroup
1 - Rural Residents	1a - Rural Retirement	1a1 - Early Rural Retirement
		1a2 - Late Stage Rural Retirement
	1b - Farming Communities	1b1 - Agricultural Communities
		1b2 - Older Farming Communities
		1b3 - Rural Commuters
	1c - Country Life	1c1 - Ageing Rural Life
		1c2 - Social Rented Rural Housing
		1c3 - Young Commuters
2 - Cosmopolitans	2a - Aspirational Migrants	2a1 - Migrant Commuters
		2a2 - Migrant Families
		2a3 - Financial Workers
	2b - Student Communities	2b1 - Student Communal Living
		2b2 - Student Digs
	2c - Settled City Living	2c1 - Older Traditional Employment
		2c2 - Established EU Service Workers
		2c3 - Urban Cultural Mix
3 - Ethnic Mix	3a - Urban Deprivation	3a1 - Striving Service Workers
		3a2 - Ageing Unemployed
	3b - Connected Achievers	3b1 - Ageing Workers
		3b2 - Multi-Ethnic Workers
		3b3 - IT Workers
	3c - Aspirational Multicultural Families	3c1 - White-Ethnic Families
		3c2 - Multi-Ethnic Families
	3d - Challenged Ethnic Mix	3d1 - Bangladeshi Hardship
		3d2 - Black Hardship
		3d3 - White Hardship
4 - Blue Collar Neighbourhoods	4a - Blue Collar Estates	4a1 - Multi-Ethnic Estates
		4a2 - Secondary Industry Workers
		4a3 - Flats and Terraced Living
	4b - Blue Collar Transitions	4b1 - Steady Transitions
		4b2 - Transitional White Neighbourhoods
		4b3 - Multi-Ethnic Industrial Transition
	4c - Blue Collar Terraces	4c1 - Senior Blue Collar Terraces
		4c2 - Ethnic Blue Collar Terraces
		4c3 - Rented Blue Collar Terraces

Supergroup	Group	Subgroup
5 - Multicultural Metropolitans	5a - Socially Mobile Minorities	5a1 - Achieving Minorities
		5a2 - Inner City Ethnic Assimilation
		5a3 - Socially Mobile New Arrivals
	5b - Ethnic Communities	5b1 - Pakistani Communities
		5b2 - Multi-Ethnic Communities
		5b3 - White-Ethnic Communities
6 - Suburbanites	6a - Inner Suburbs	6a1 - Multi-Ethnic Suburbs
		6a2 - Young Workers in Terraces
		6a3 - Ageing Europeans
	6b - Established Suburbs	6b1 - Ageing in Suburbia
		6b2 - Young Suburban Family Tenants
		6b3 - Suburban Service Workers with Higher Qualifications
	6c - Suburban Aspiration	6c1 - Detached Retirement Living
		6c2 - Semi-Detached Suburbia
		6c3 - Young Families in Detached Houses
7 - Hard-Pressed Households	7a - Industrial Legacy	7a1 - Young Hard-Pressed Families
		7a2 - Old Industrial Workers
	7b - Hard-Pressed Multi-Ethnic Neighbourhoods	7b1 - Hard-Pressed Adults with Prospects
		7b2 - Hard-Pressed European Settlers
		7b3 - Deprived and Isolated Ethnic Minorities
	7c - Elderly in Flats	7c1 - Settled Hard-Pressed Pensioners
		7c2 - Dependant Hard-Pressed Pensioners
		7c3 - Deprived Elderly Communities
8 - Urbanites	8a - Traditional Trades	8a1 - Achieving Tradespeople
		8a2 - Educated Tradespeople
		8a3 - Striving Multi-Ethnic Tradespeople
	8b - Service Sector Urbanities	8b1 - Ageing Urban Service Workers
		8b2 - Young Urban Service Workers
		8b3 - Skilled Urban Service Workers
	8c - Late Retirement	8c1 - Delayed Retirement
		8c2 - Self-Sufficient Retirement
		8c3 - Communal Retirement

Appendix D

D.1. Similarities of each OA and SA in the UK to the 2011 OAC Supergroups

Section D.1 contains choropleth maps that visualise the propensity for each OA and SA in the UK to conform to each of the eight 2011 OAC Supergroups.

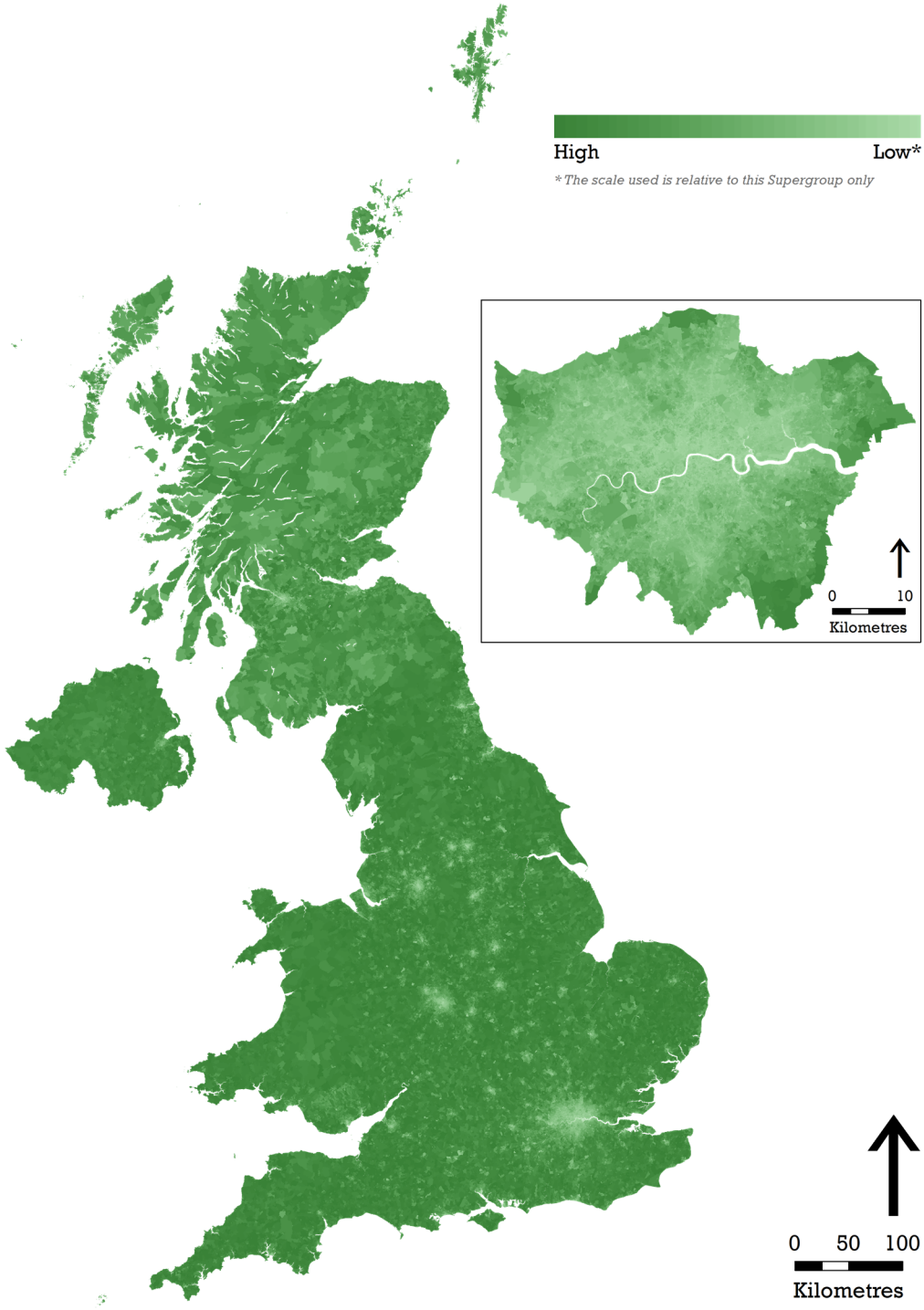


Figure D.1: The similarities of each OA and SA in the UK to the 'Rural Residents' 2011 OAC Supergroup

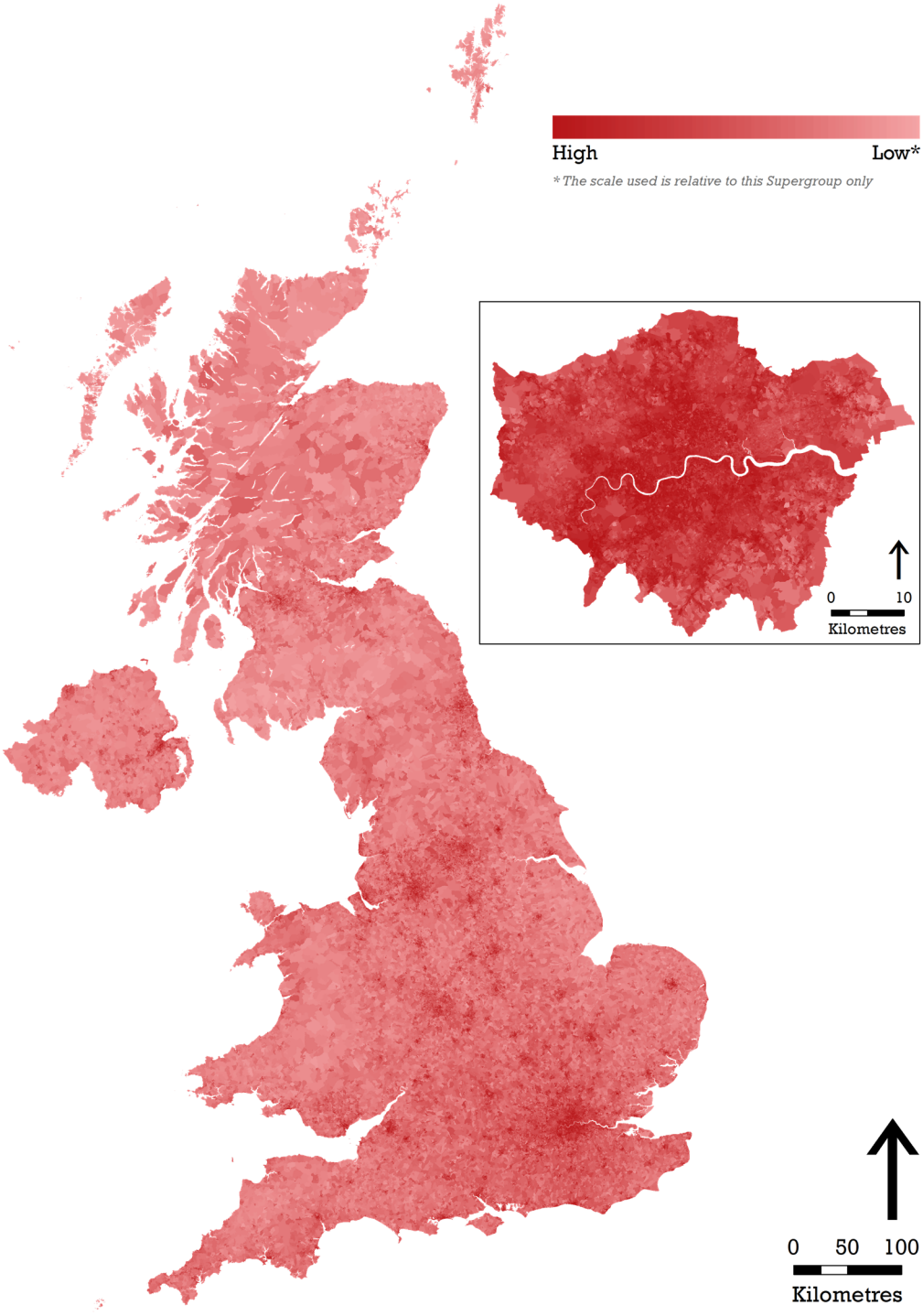


Figure D.2: The similarities of each OA and SA in the UK to the 'Cosmopolitans' 2011 OAC Supergroup

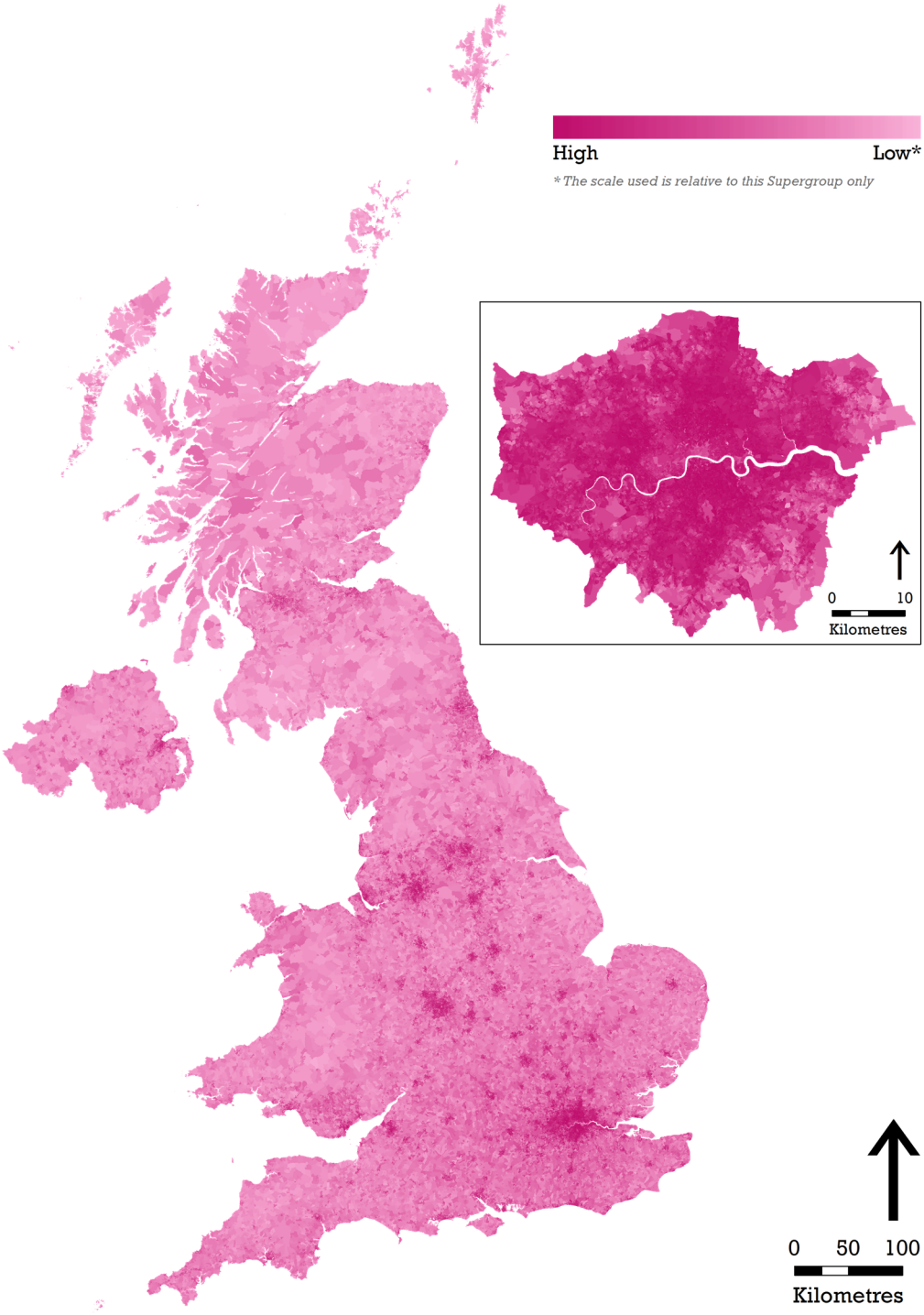


Figure D.3: The similarities of each OA and SA in the UK to the 'Ethnicity Central' 2011 OAC Supergroup

Figure D.4: The similarities of each OA and SA in the UK to the 'Multicultural Metropolitans' 2011 OAC Supergroup

Figure D.5: The similarities of each OA and SA in the UK to the 'Urbanites' 2011 OAC Supergroup

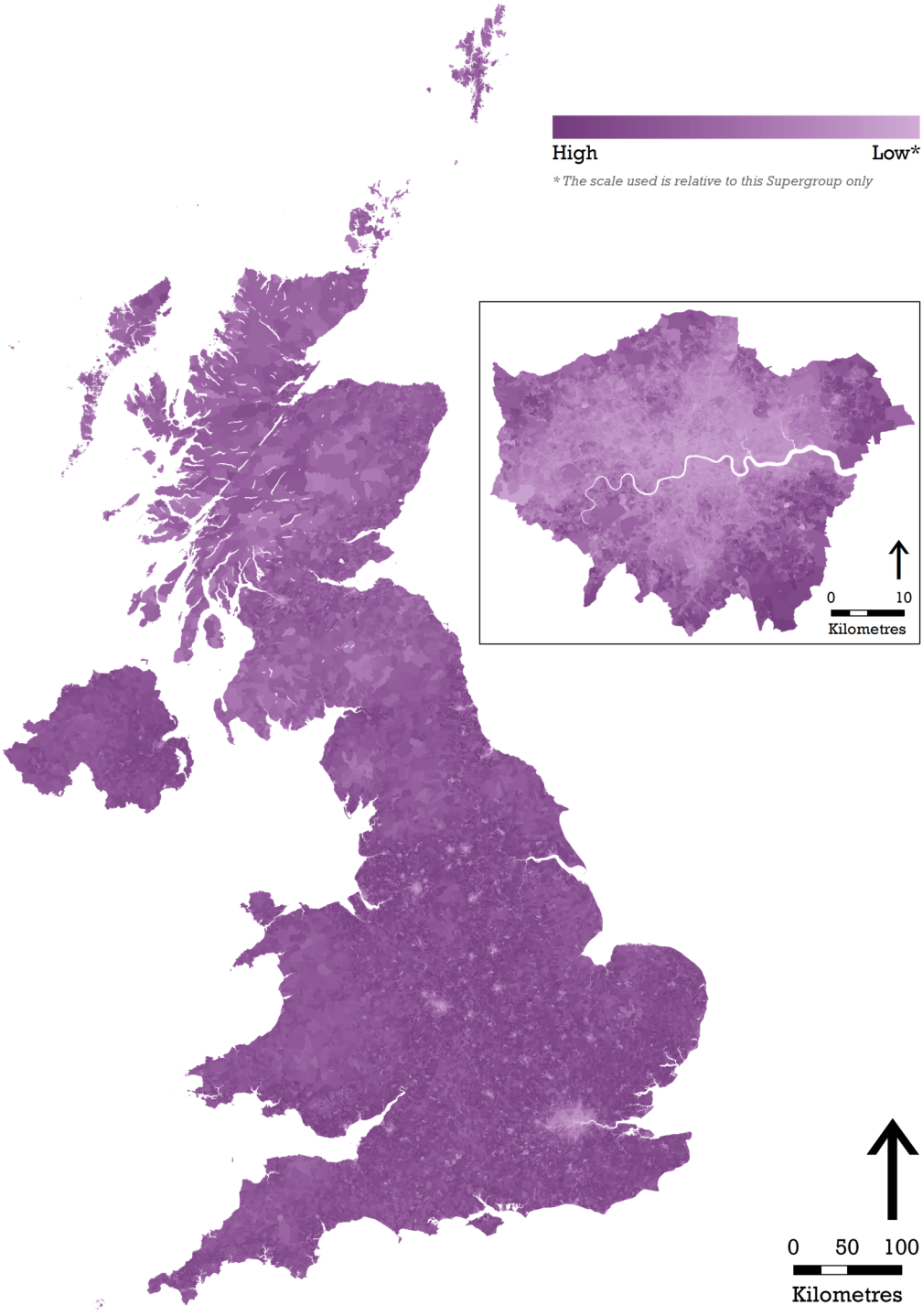


Figure D.6: The similarities of each OA and SA in the UK to the 'Suburbanites' 2011 OAC Supergroup

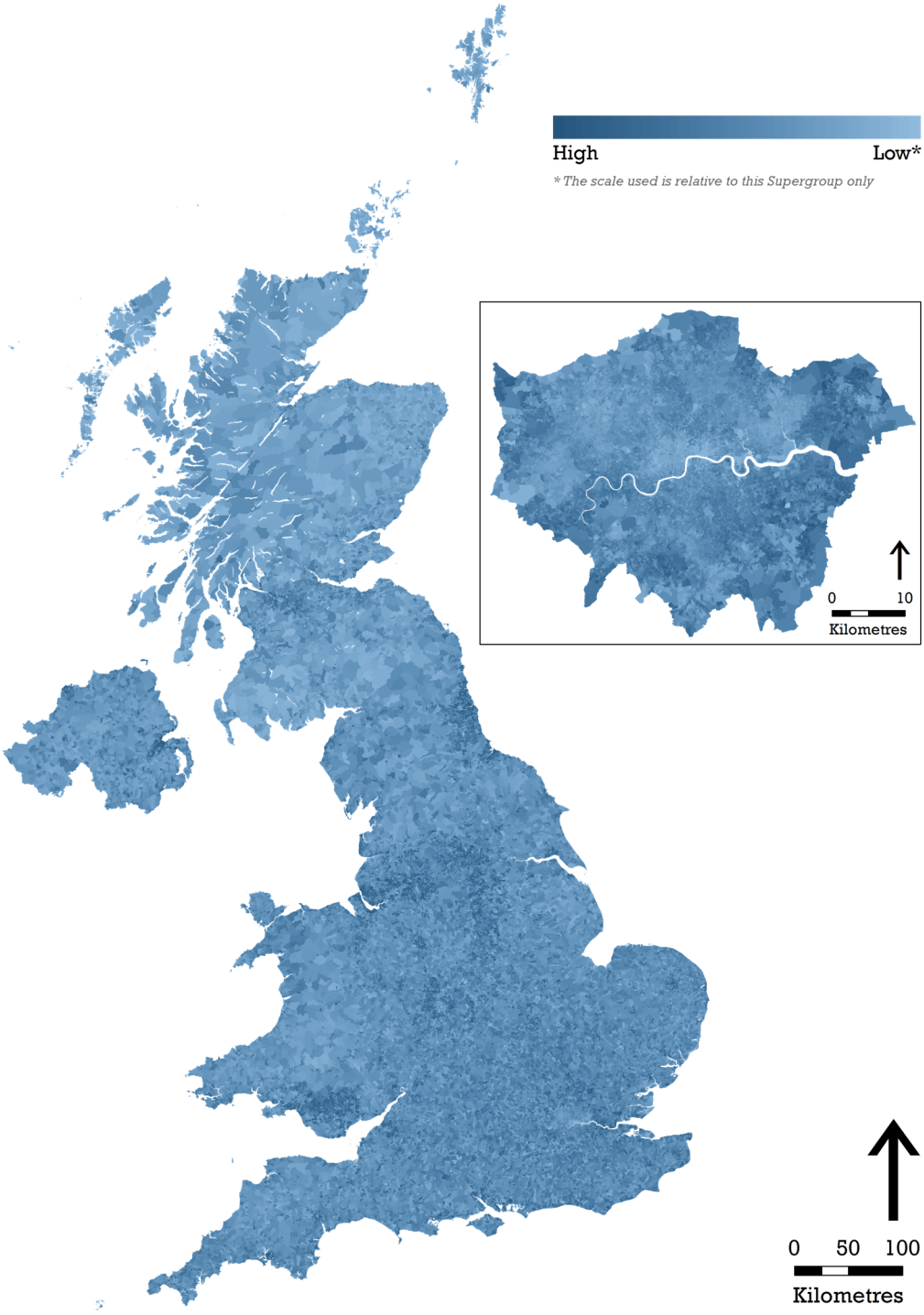


Figure D.7: The similarities of each OA and SA in the UK to the 'Constrained City Dwellers' 2011 OAC Supergroup

Figure D.8: The similarities of each OA and SA in the UK to the 'Hard-Pressed Living' 2011 OAC Supergroup

D.2. The Gini Coefficients of the 2011 OAC variables

Table D.1: Gini Coefficients of the 60 variables used to construct the 2011 OAC

Code	Variable Name	Gini Coefficient
k001	Persons aged 0 to 4	0.134
k002	Persons aged 5 to 14	0.093
k003	Persons aged 25 to 44	0.051
k004	Persons aged 45 to 64	0.045
k005	Persons aged 65 to 89	0.118
k006	Persons aged 90 and over	0.596
k007	Number of persons per hectare	0.184
k008	Persons living in a communal establishment	0.911
k009	Persons aged over 16 who are single	0.074
k010	Persons aged over 16 who are married or in a registered same-sex civil partnership	0.047
k011	Persons aged over 16 who are divorced or separated	0.080
k012	Persons who are white	0.021
k013	Persons who have mixed ethnicity or are from multiple ethnic groups	0.486
k014	Persons who are Asian/Asian British: Indian	0.688
k015	Persons who are Asian/Asian British: Pakistani	0.808
k016	Persons who are Asian/Asian British: Bangladeshi	0.888
k017	Persons who are Asian/Asian British: Chinese and Other	0.598
k018	Persons who are Black/African/Caribbean/Black British	0.675
k019	Persons who are Arab or from other ethnic groups	0.750
k020	Persons whose country of birth is the United Kingdom or Ireland	0.029
k021	Persons whose country of birth is in the old EU (pre 2004 accession countries)	0.451
k022	Persons whose country of birth is in the new EU (post 2004 accession countries)	0.555
k023	Main language is not English and cannot speak English well or at all	0.613
k024	Households with no children	0.075
k025	Households with non-dependant children	0.120
k026	Households with full-time students	0.903
k027	Households who live in a detached house or bungalow	0.297
k028	Households who live in a semi-detached house or bungalow	0.183
k029	Households who live in a terrace or end-terrace house	0.265
k030	Households who live in a flat	0.362
k031	Households who own or have shared ownership of property	0.061

Code	Variable Name	Gini Coefficient
k032	Households who are social renting	0.347
k033	Households who are private renting	0.176
k034	Occupancy room rating -1 or less	0.289
k035	Individuals day-to-day activities limited a lot or a little Standardised Illness Ratio	0.047
k036	Persons providing unpaid care	0.073
k037	Persons aged over 16 whose highest level of qualification is Level 1, Level 2 or Apprenticeship	0.042
k038	Persons aged over 16 whose highest level of qualification is Level 3 qualifications	0.066
k039	Persons aged over 16 whose highest level of qualification is Level 4 qualifications and above	0.088
k040	Persons aged over 16 who are schoolchildren or full-time students	0.141
k041	Households with two or more cars or vans	0.116
k042	Persons aged between 16 and 74 who use public transport to get to work	0.202
k043	Persons aged between 16 and 74 who use private transport to get to work	0.059
k044	Persons aged between 16 and 74 who walk, cycle or use an alternative method to get to work	0.143
k045	Persons aged between 16 and 74 who are unemployed	0.209
k046	Employed persons aged between 16 and 74 who work part-time	0.037
k047	Employed persons aged between 16 and 74 who work full-time	0.012
k048	Employed persons aged between 16 and 74 who work in the agriculture, forestry or fishing industries	0.764
k049	Employed persons aged between 16 and 74 who work in the mining, quarrying or construction industries	0.122
k050	Employed persons aged between 16 and 74 who work in the manufacturing industry	0.152
k051	Employed persons aged between 16 and 74 who work in the energy, water or air conditioning supply industries	0.463
k052	Employed persons aged between 16 and 74 who work in the wholesale and retail trade; repair of motor vehicles and motor cycles industries	0.062
k053	Employed persons aged between 16 and 74 who work in the transport or storage industries	0.190
k054	Employed persons aged between 16 and 74 who work in the accommodation or food service activities industries	0.178
k055	Employed persons aged between 16 and 74 who work in the information and communication or professional, scientific and technical activities industries	0.159
k056	Employed persons aged between 16 and 74 who work in the financial, insurance or real estate industries	0.211
k057	Employed persons aged between 16 and 74 who work in the administrative or support service activities industries	0.173
k058	Employed persons aged between 16 and 74 who work in the in public administration or defence; compulsory social security industries	0.164
k059	Employed persons aged between 16 and 74 who work in the education sector	0.110
k060	Employed persons aged between 16 and 74 who work in the human health and social work activities industries	0.069

Appendix E

E.1. Published Journal Papers

Gale, C. G. and Longley, P. A. (2013), Temporal Uncertainty in a Small Area Open Geodemographic Classification. Transactions in GIS, 17: 563–588. doi: 10.1111/tgis.12035